Bachelorarbeit

Institut für Statistik
Ludwig-Maximilians-Universität München

Versuchsplanung und Parameteroptimierung für geophysikalische Computersimulationsstudien

Paul Messer

Projektpartner:
Sebastian Anger
Department für Geo- und Umweltwissenschaften (Geophysik)
LMU München

Betreuer:
Dr. Fabian Scheipl
Datum: 14. November 2019

Zusammenfassung

In Zusammenarbeit mit dem Department für Geophysik der Ludwig-Maximilians-Universität wurde in dieser Bachelor Thesis untersucht, welche numerischen Parameter für gegebene physikalische Parameter bei einer Erdbebensimulation möglichst stabile und gleichzeitig, im Sinne von Rechenzeit, kostengünstige Ergebnisse liefern. Im Bereich der Erdbebenphysik werden häufig numerische Simulationen verwendet, um physikalische Prozesse realistisch darstellen zu können. Hier wird der relative Unterschied zwischen einer numerischen und der exakten Lösung bezüglich des Porendrucks betrachtet, jener sollte möglichst minimal sein. In dieser Parameterstudie wird eine Methode gesucht, welche für gegebene physikalische Parameter und einen gegebenen maximalen relativen Fehler, numerische Parameter ermittelt. Diese sollten gleichzeitig auch möglichst kostengünstig im Sinne von Rechenintensität sein. Um die Datengrundlage zu erzeugen, welche das Verhalten des relativen Fehlers zwischen exakter und numerischer Lösung, abhängig von physikalischen und numerischen Parametern, beinhaltet, wurde ein *Latin hypercube sampling* angewandt. Dadurch wird der Parameterraum möglichst optimal abgedeckt. Dieser Versuchsplan wurde nach der Betrachtung seiner Ergebnisse in bestimmten Parameterbereichen verfeinert, um in diesen Bereichen genauere Aussagen über den Einfluss der Parameter treffen zu können. Um die Rechenzeit dabei zu berücksichtigen, wurde der Einfluss der numerischen Parameter auf die Rechenzeit betrachtet. Die Rechenzeit lässt sich durch den numerischen Parameter „Anzahl der Gridpunkte“ erklären. Um den Einfluss der verschiedenen Parameter auf die Abweichung der numerischen zur exakten Lösung bezüglich des Porendrucks zu bestimmen, wurde ein Regressionsmodell entwickelt. Dieses wird anschließend genutzt, um für gegebene physikalische Parameter die erwartete niedrigste Anzahl der Gridpunkte zu bestimmen, welche benötigt wird, um unter der gegebenen Fehlerschranke zu bleiben. Jenes Regressionsmodell wurde schlussendlich invertiert, um für gegebene physikalische Parameter und eben jene Fehlerschranke eine geeignete Schätzung der für die Simulation notwendigen Anzahl an Gridpunkten zu liefern.

Inhaltsverzeichnis

1 Einleitung

In Zusammenarbeit mit dem Department für Geophysik der Ludwig-Maximilians-Universität wurde in dieser Abschlussarbeit untersucht welche numerischen Parameter für gegebene physikalische Parameter bei einer Erdbebensimulation möglichst stabil und gleichzeitig, im Sinne von Rechenzeit, kostengünstig sind. Die Erdbebenmagnitude spielt im Bereich der Erdbebenphysik eine zentrale Rolle. Die Magnitude ist ein Maß, welches die Stärke von Erdbeben angibt. Je stärker ein Erdbeben ist, desto verheerender die Auswirkungen. Um die Gefahren und Ausbreitungen von Erdbeben besser ermitteln zu können, muss man die physikalischen Prozesse analysieren und verstehen. Dafür werden häufig numerische Simulationen verwendet, welche die physikalischen Prozesse realistisch darstellen sollen. Zu diesem Zweck wurde hier ein physikalischer Prozess in einem numerischen Tool implementiert. Es geht darum zu untersuchen, welchen Einfluss der Porendruck während eines Erdbebens hat. Des Weiteren benötigt man für die Implementierung weitere numerische Parameter, damit die Simulation stabil bleibt und sich einer analytischen Lösung annähert. Dementsprechend ist man immer darauf bedacht, dass der relative Fehler zwischen exakter und numerischer Lösung bezüglich des Porendrucks minimal und dabei auch kostengünstig im Sinne von Rechenintensivität ist. Der relative Fehler zwischen exakter und numerischer Lösung sollte die Fehlerschranke 1.00E−06 nicht überschreiten. Die physikalischen Einflussgrößen sind hierbei die thermische Diffusivität, die hydraulische Diffusivität, die spezifische Wärmekapazizät, die Druckänderung pro Temperaturanstieg, die Breite der Bruchzone und die Bruchzonenstärke. Bei Simulationen von Erdbeben werden an verschiedenen Gridpunkten, welche hinsichtlich ihrer Anzahl und Räumlichkeit variieren, die numerischen Werte berechnet. Für eine stabile und rechenarme Berechnung gilt es möglichst präzise zu quantifizieren, wie sich die numerischen Parameter verhalten. Diese sind der Gridabstand, die Anzahl der Gridpunkte und die minimale sowie maximale Länge, um die Diffusionsgleichung zu lösen. Im Kapitel über die Datenbasis (Abschnitt 2) dieser Abschlussarbeit wird erklärt, welche Ausprägungen die physikalischen Parameter annehmen können und wie jene zusammenhängen. Danach wird die methodische Grundlage erläutert, auf welcher im Auswertungskapitel (Abschnitt 4) die Problemstellung statistisch untersucht, modelliert und bewertet wird.

1.1 Software

Für alle hier getätigten Analysen wurde die Statistik-Software R (Version 3.6.1, R Core Team; 2019) verwendet. Dabei wurden verschiedene R-Packages genutzt. Ein

Großteil der Visualisierungen wurden mit dem `ggplot2`-Package (Wickham; 2016) und dem `gridExtra`-Package (Auguie; 2017) gemacht. Die Visualisierung der partiellen Residuen wurde mittels des `car`-Packages (Fox and Weisberg; 2019) erstellt. Das Erstellen der *Latin Hypercube Samplings* in Abschnitt 2.3 erfolgte über das `tgp`-Package (Gramacy and Taddy; 2010). Die Berechnung des RMSE in Abschnitt 3.3 erfolgte durch das `caret`-Package (from Jed Wing et al.; 2019). Dabei wurden die genutzten *Kreuzvalidierungen* durch das `cvTools`-Package (Alfons; 2012) durchgeführt. Die B-Splines wurden mittels des `splines`-Packages (R Core Team; 2019) berechnet. Abschließend erfolgte die Inversion des Modells durch das `investr`-Package (Greenwell and Kabban; 2014).

2 Datenbasis

2.1 Physikalische Parameter

Die physikalischen sowie die numerischen Parameter haben bestimmte Ober- und Untergrenzen. Der Parameter „der spezifischen Wärmekapazität“ (rho_c) wird auf $rho_c = 2.70\text{E+}06\frac{Pa}{K}$ festgelegt. Die möglichen Intervalle der Parameter sind in Tabelle 1 aufgeführt.

physikalische Parameter	Variablenname	Minimum	Maximum	Einheit
thermische Diffusivität	α_{th}	5.40E−07	1.00E+06	$\frac{m2}{s}$
hydraulische Diffusivität	α_{hy}	8.60E−07	7.15E+06	$\frac{m2}{s}$
Druckänderung pro Temperaturanstieg	λ	6.80E+04	9.80E+05	$\frac{Pa}{K}$
Breite der Bruchzone	hwid	2.50E−05	1.00E−02	m

Tabelle 1: Physikalische Parameterwerte

2.2 Numerische Parameter

Der numerische Parameter „Rasterabstand“ ($dlDwn$) hängt von den numerischen Parametern, welche den minimalen (Dwn_{min}) und den maximalen (Dwn_{max}) Wert der Länge der Diffusion angeben, ab. Jene wiederum hängen von dem physikalischen Parameter „Breite der Bruchzone“ („hwid“) ab. Die numerischen Parameter sind, ausgenommen von der Anzahl an Gridpunkten, stetig. Der „Rasterabstand“ ($dlDwn$) ist definiert als $\frac{ln(Dwn_{min}/Dwn_{max})}{nz-1}$. $\frac{10}{hwid}$ definiert den maximalen Wert der Länge der Diffusion.

numerische Parameter	Variablenname	Minimum	Maximum
minimaler Wert der Länge der Diffusion	Dwn_{min}	$\frac{1.00\text{E}-16}{hwid}$	$\frac{1}{hwid}$
Anzahl der Gridpunkte	nz	10	1000

Tabelle 2: Numerische Parameterwerte

2.3 Versuchsplan

Für die Abdeckung des ganzen Parameterraums wurde ein mittels eines *Latin Hypercube Designs* erstellter Versuchsplan genutzt. Auf den genaueren theoretischen Hintergrund wird in Abschnitt 3, welcher die methodischen Grundlagen beinhaltet, genauer eingegangen. Für Bereiche, in denen kleinere Abstände zwischen den Simulationsparametern benötigt wurden, sind jeweils *Latin Hypercube Designs* mit veränderten Minima und Maxima für die jeweiligen Parameter gezogen worden. Dadurch sind die Daten auf der jeweiligen X-Achse nicht gleichverteilt. Bei der Visualisierung des Einflusses dieser Variablen wird ein entsprechendes Gitter verwendet, welches die verschiedenen Achsenabschnitte abdeckt.

2.4 Deskription

Hier wird die Verteilung der jeweiligen physikalischen Parameter visualisiert, welche auf ihrem Intervall nicht gleichverteilt sind, da sie, wie in Kapitel 2.3 beschrieben, in kleineren Bereichen der Intervalle feiner abgetastet wurden. Daraus folgen die in Abbildung 1 gezeigten Verteilungen auf dem jeweiligen Parameterintervall.

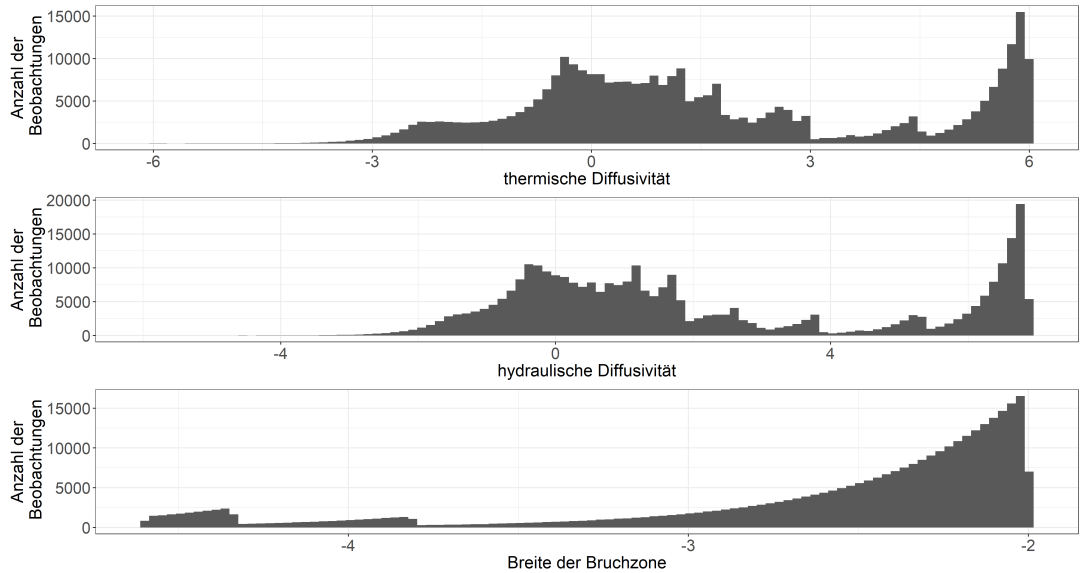

Abbildung 1: Die Verteilung der physikalischen Parameter „thermische Diffusivität“, „hydraulische Diffusivität“ und „Breite der Bruchzone“

3 Methodik

Im Rahmen dieser Abschlussarbeit werden verschiedene statistische Methoden angewandt, um sich der Problemstellung der Paramateroptimierung für eine gegebene Fehlerschranke anzunehmen. In diesem Kapitel wird die theoretische Grundlage beschrieben, auf welcher die Problemstellung im folgenden Kapitel praktisch untersucht wird. Zum einen wird erklärt, auf welcher Grundlage der Versuchsplan erstellt wurde und welche Theorie dahinter steht. Zum anderen wird auf die Modellierung einer Regression eingegangen, es werden die dafür notwendigen Voraussetzungen definiert und die daraus resultierende Schätzgenauigkeit wird erläutert. Darüber hinaus werden verschiedene Modellierungsansätze erläutert und methodisch erklärt, wie jene bezüglich ihrer Güte zu vergleichen sind. Schlussendlich wird ausgeführt, wie gefundene Modelle invertiert werden können und was in Bezug auf die Schätzgenauigkeit der Inversion zu beachten ist.

3.1 Latin Hypercube Sampling

Um den Parameterraum gleichmäßig abzudecken, wurde hier der Versuchsplan über ein *Latin Hypercube Design* (*LHD*) erstellt. Ein *Latin Hypercube Design* besteht aus einer $n \times n_r$-Matrix, bei welcher jede Spalte $r = 1, 2, 3, ..., n_r$ aus einer zufälligen Permutation besteht. Um daraus ein *Latin Hypercube Sampling* (*LHS*) zu erzeugen, wird jedem Wert des *LHD* eine Zufallszahl aus dem halboffenen Intervall $[0, 1)$ abgezogen und dann durch die Anzahl an Werten (n) geteilt. Somit entsteht ein Testfeld im Einheitsraum (jede Spalte hat Werte im Intervall [0,1]). Der Einheitsraum entsteht, da die abgezogenen Werte im halboffenen Intervall $[0, 1)$ liegen. Um für diese Parameterstudie die Werte nun an die möglichen Ausprägungen der jeweiligen Simulationsparameter anzupassen, werden diese mit der Spannweite multipliziert und zu dem Minimum der möglichen Werte addiert.

$$x_{(ij)} = \frac{x_{(ij)}^{LHD} - Zufallszahl[0,1)}{n_r}, \quad \text{mit } x_{(ij)}^{LHD} \in 1, 2, 3, ..., n \tag{1}$$

Eine Visualisierung dessen sieht man in Abbildung 2. Man erkennt, dass den Punkten im *LHS* jeweils ein zufälliger Wert in X- sowie in Y-Richtung abgezogen wurde. Jedoch ergibt sich aus der Konstruktion eines *LHS* kein garantiert gleichverteiltes und korrelationsfreies Testfeld. Dies erübrigt sich hierbei jedoch durch die große Anzahl an gezogenen Punkten. Hier wurden aufgrund des hohen Rechenaufwands

an vielen verschiedenen Simulationscomputern *Latin Hypercube Samplings* erstellt und insgesamt über 300.000 Datenpunkte durch das numerische Tool generiert. Da ein *LHS* die Simulationsparameter auf dem Parameterraum grundsätzlich annähernd gleichverteilt, wurden insbesondere die Simulationsparameter mit sehr großen Spannweiten, aber gleichzeitig kleinen Minima, nicht optimal abgedeckt, um über extrem kleine Parameterausprägungen Aussagen treffen zu können. Somit wurden weitere *Latin Hypercube Samplings* gezogen, bei denen der Einheitsraum auf ein Intervall mit niedrigerem Maximum zurückgeführt wurde. Wenn man beispielsweise den Parameter α_{th} betrachtet, gehen dessen mögliche Ausprägungen von kleinen Bereichen nahe Null (5.40E−07) bis hin zu Werten fernab Null (1.00E+06). Nun wäre es durch ein einfaches *LHS* in diesem Intervall kaum möglich Aussagen über die Bereiche zwischen $[5.40\text{E}{-}07, 1]$ zu treffen. Dies wurde durch die geänderten Maxima angepasst, wodurch nun aber die Parameter nicht mehr annähernd gleichverteilt auf ihrem Intervall liegen (siehe Abschnitt 2.3). Ein *LHS* hat, gegenüber einem rein zufälligen *Monte-Carlo-Feld*, den Vorteil, dass die Varianz des Mittelwerts geringer ist (Stein; 1987, S. 143). Eine beispielhafte Verteilung eines zweidimensionalen *LHD* bzw. *LHS* ist nachfolgender Abbildung 2 zu entnehmen.

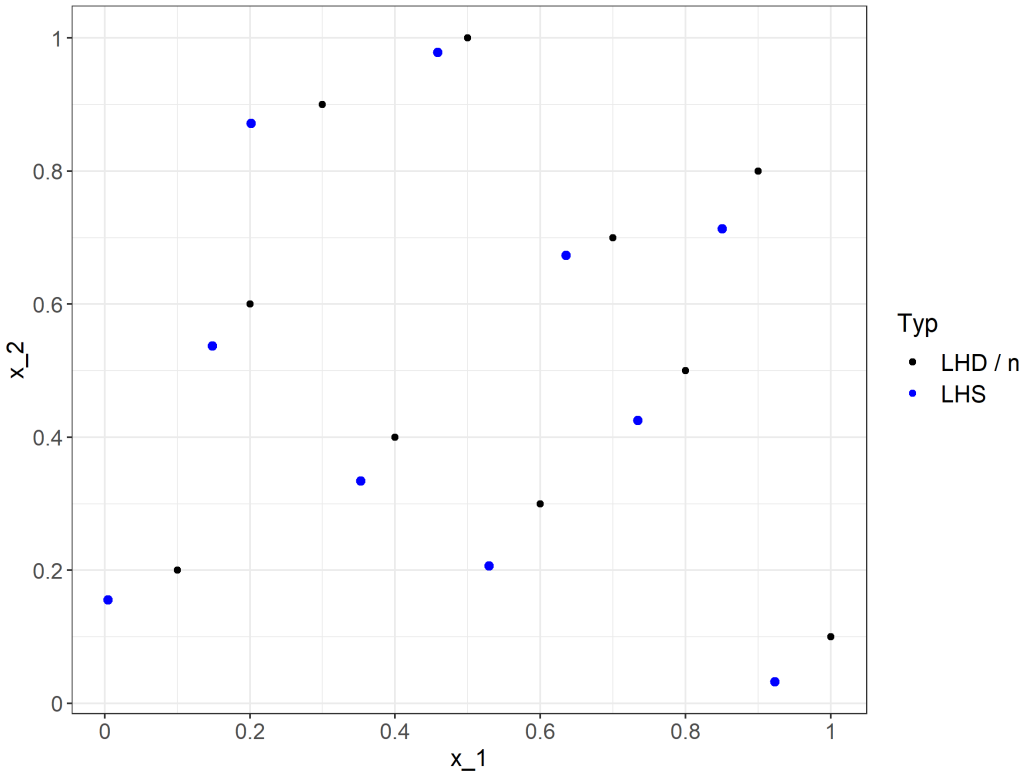

Abbildung 2: Beispielverteilung eines zweidimensionalen *Latin Hypercube Designs* (*LHD*) bzw. *Latin Hypercube Samplings* (*LHS*) für die Variablen x_1 und x_2

3.2 Modellierung

3.2.1 Lineares Modell

Um den Zusammenhang zwischen der Abweichung der numerischen Lösung zur exakten Lösung bezüglich des Porendrucks abhängig von den Simulationsparametern modellieren zu können, wird eine Regressionsanalyse angewandt. Eine Regressionsanalyse versucht, eine Zielvariable y in Abhängigkeit von Kovariablen (Regressoren) $x_1, x_2, ..., x_k$ zu erklären. Entsprechend wird ein Erwartungswert $E(y)$ abhängig von den Regressoren modelliert. Zusätzlich wird ein Störterm ϵ aufgenommen, welcher die zufällige, nicht von Kovariablen erklärte, Abweichung vom Erwartungswert berücksichtigt (Fahrmeir et al.; 2009, S. 19). Eine Visualisierung dessen befindet sich in Abbildung 3. Daraus ergibt sich die folgende Zerlegung der Zielvariable y:

$$y = E(y|x_1, ..., x_n) + \epsilon = f(x_1, ..., x_n) + \epsilon \tag{2}$$

Im Falle eines linearen Regressionsmodells wird der Funktion f unterstellt, dass sie linear ist, wodurch sich:

$$\hat{y} = E(y|x_1, ..., x_n) + \epsilon = \beta_0 + \beta_1 x1 + ... + \beta_k x_k \tag{3}$$

ergibt. $\beta_0, \beta_1, ..., \beta_k$ stehen hierbei für die Regressionskoeffizienten des jeweiligen x_i. β_0 ist der sogenannte Intercept, welcher bei ausschließlich metrischen Einflussvariablen jenes Y angibt, welches für alle $x_i = 0$ geschätzt wird. Im Anwendungsfall dieser Abschlussarbeit sieht ein rein lineares Modell beispielsweise folgendermaßen aus:

$$\widehat{error}_p = \beta_0 + \beta_{\alpha_{th}}\alpha_{th} + \beta_{\alpha_{hy}}\alpha_{hy} + \beta_\lambda \lambda + \beta_{hwid} hwid + \beta_{nz} nz \tag{4}$$

Für das Modell aus Gleichung 2 müssen die Annahmen gelten, dass die Fehler $\epsilon_1, ..., \epsilon_k$ unabhängig und identisch verteilt sind; sie haben den Erwartungswert 0: $E(\epsilon_i) = 0$. Oftmals wird zusätzlich eine Normalverteilung mit Erwartungswert 0: $E(\epsilon_i) = 0$ und Varianz $Var(\epsilon_i) = \sigma^2$ angenommen. Die Varianz des Störterms ist konstant über die Beobachtungen (Homoskedastizität). Zusätzlich gilt die Annahme, dass die Störterme unkorreliert sind: Für $i \neq j$ gilt: $Cov(\epsilon_i, \epsilon_j) = 0$. Im Folgenden werden die Komponenten der Regression in Vektoren- bzw. Matrixschreibweise dar-

gestellt und es gelten folgende Definitionen:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_n \\ \vdots \\ \epsilon_i \end{pmatrix}$$
$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & ... & x_1 k \\ 1 & \vdots & \vdots & \vdots \\ 1 & x_{n1} & ... & x_{nk} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x_1'} \\ \vdots \\ \boldsymbol{x_n'} \end{pmatrix} \tag{5}$$

Die Gleichung 2 lässt sich somit kompakter in Matrixschreibweise schreiben:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{6}$$

β_i ist hier ein Koeffizient, welcher als Steigungsparameter bezeichnet werden kann, der angibt, wie sich die Zielvariable y verändert, wenn das zugehörige x_i um eine Einheit steigt. Der jeweilige Parameter β_i wird über die Methode der kleinsten Quadrate geschätzt. Der kleinste Quadrate Schätzer (KQ-Schätzer) bestimmt die jeweiligen β_i, indem er die Summe der quadratischen Abweichung bei gegebenen Daten minimiert. Die geschätzten Werte $\hat{y}$ des KQ-Schätzers sind orthogonal zu den Residuen $\hat{\epsilon}$ (Fahrmeir et al.; 2009, S. 97).

Für den KQ-Schätzer gilt:

$$KQ(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - x_i'\beta)^2 = \boldsymbol{\epsilon'}\boldsymbol{\epsilon} \tag{7}$$

Das Minimierungsproblem wird durch das Nullsetzen der Ableitung gelöst und daraus folgt der KQ-Schätzer (Fahrmeir et al.; 2009, S. 92):

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y} \tag{8}$$

Das ˆ Symbol ist die gängige Notation dafür, dass es sich um eine Schätzung handelt und wird im Folgenden weiterverwendet. Zusätzlich zur Modellierung linearer Effekte werden im Rahmen dieser Abschlussarbeit noch Interaktionseffekte aufgenommen. Interaktionseffekte bezeichnen nicht-additive Effekte zweier oder mehrerer Regressoren. Interpretiert werden kann dies beispielsweise als: Die Wirkung von α_{th} ist abhängig vom Wert der Variable α_{hy}. Umgesetzt wird dies, indem die Variablen jeweils miteinander multipliziert werden und somit eine neue Einflussvariable bilden, jene wird dann wie die anderen Regressoren behandelt. Somit ergibt sich bei einer

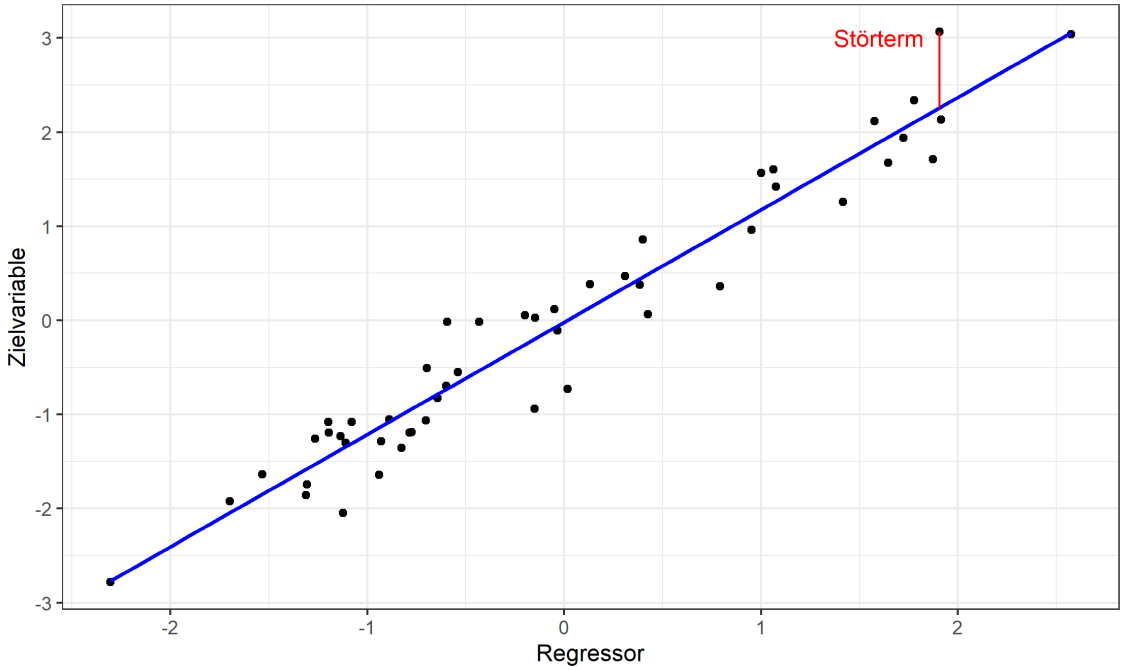

Abbildung 3: Eine Regressionsgerade, basierend auf einem Beispieldatensatz, welche den quadrierten Abstand der Punkte zur Geraden minimiert. Die ϵ_i sind hier die Abstände zwischen der Gerade und den jeweiligen Punkten. Ein ϵ wurde beispielsweise am rechten oberen Rand des Plots visualisiert.

möglichen Interaktion zwischen den Variablen α_{th} und α_{hy} die Variable $inter_{th_hy}$, welche definiert ist als: $inter_{th_hy} = \alpha_{th} \cdot \alpha_{hy}$

Das Modell 4 wird damit beispielsweise folgendermaßen erweitert:

$$\begin{aligned} error_p = \beta_0 + \beta_{\alpha_{th}}\alpha_{th} + \beta_{\alpha_{hy}}\alpha_{hy} + \beta_\lambda\lambda + \\ \beta_{hwid}hwid + \beta_{nz}nz + \beta_{inter_{thhy}}inter_{thhy} + \epsilon \end{aligned} \tag{9}$$

Im weiteren Verlauf dieser Abschlussarbeit (Kapitel 3.3) wird anhand der Wurzel der mittleren Fehlerquadratsumme (bzw.: des Root-mean-squared errors, kurz: RMSE) verglichen, welche Interaktionen das Modell verbessern und schlussendlich in ein finales Modell aufgenommen werden. Zusätzlich dazu werden die partiellen Residuen $\epsilon_{x_j,i}$ betrachtet. Die partiellen Residuen sind, mit Ausnahme vom betrachteten x_j, um den Einfluss aller Kovariablen bereinigt und folgendermaßen definiert (Fahrmeir et al.; 2009, S. 110):

$$\hat{\epsilon}_{xj,i} = y_i - \hat{\beta}_0 - ... - \hat{\beta}_{j-1}x_{i,j-1} - \hat{\beta}_{j+1}x_{i,j+1} - ... - \hat{\beta}_k x_{i,k} = \hat{\epsilon}_i + \hat{\beta}_j x_{i,j} \tag{10}$$

Die partiellen Residuen eignen sich dafür, die gewählte Modellierung des Einflusses von x_j zu überprüfen.

3.2.2 B-Splines

Um einen nichtlinearen Einfluss einer metrischen Einflussvariable auf eine wiederum metrische Zielvariable flexibler zu modellieren, werden in der Statistik oft sogenannte Splines genutzt. Gesucht wird also ein Modell mit einer Funktion, welche den Erwartungswert der Zielvariable y möglichst präzise modellieren kann:

$$\begin{aligned} y_i &= f(z_i) + \epsilon_i \\ f &: [a, b] \rightarrow R \end{aligned} \tag{11}$$

f ist dabei ein Polynom-Spline vom Grad $l \geq 0$ zu den Knoten $a = k_1 < ... < k_m = b$. Für diese Funktion $f(z)$ gilt, dass sie $l - 1$ mal stetig differenzierbar sein muss. Ein Sonderfall ist wenn $l = 0$ gilt. In diesem Fall wäre keine Glattheitsanforderung an $f(z)$ gestellt. $f(z)$ muss auf den durch die Knoten gebildeten Intervallen $[k_j, k_{j+1}]$ ein Polynom vom Grad l sein (Fahrmeir et al.; 2009, S. 295). Für dieses Modell gelten außerdem $E(\epsilon_i) = 0$ und $Var(\epsilon_i) = \sigma^2$.
Im Rahmen dieser Abschlussarbeit werden Basic-Splines (B-Splines) genutzt, die anderen Splines (z.B.: P-Splines auf TP-Basis) gegenüber insbesondere aus numerischer Sicht vorzuziehen sind (Fahrmeir et al.; 2009, S. 303). Splines mit einer TP-Basis haben den praktischen Nachteil, dass sie zu Kollinearität neigen (Ruppert et al.; 2003, S. 70). B-Splines haben den Vorteil, dass die Anpassung von Daten durch den KQ-Schätzer zu einem linearen Problem wird (Blobel and Lohrmann; 2013, S. 231). Ein B-Spline ist eine Funktion, welche stückweise aus abgeschnittenen Polynomen desselben Grades zusammengesetzt ist. Damit B-Splines die notwendigen Glattheitsanforderungen an die Funktion $f(z)$ an den Knoten erfüllen, werden die Basisfunktionen auf die Art und Weise konstruiert, dass Polynomstücke des gewünschten Grades ausreichend glatt zusammengesetzt werden (Fahrmeir et al.; 2009, S. 303). Eine B-Spline-Basisfunktion ist aus $l + 1$ Polynomstücken vom Grad l zusammengesetzt. Die Darstellung der Basisfunktionen befindet sich für Grad $l = 0, ..., 3$ in Abbildung 4, entnommen aus (Fahrmeir et al.; 2009, S. 303).

Hierbei ist zu beachten, dass ausschließlich äquidistante, also gleichmäßig verteilte, Knoten betrachtet werden. Dabei wird für die Knoten der Wertebereich $[a, b]$ von z in $m - 1$ Intervalle mit der Breite $h = \frac{b-a}{m-1}$ eingeteilt. Bei äquidistanten Knoten haben alle Basisfunktionen dieselbe Gestalt und sind nur entlang der z-Achse verschoben.

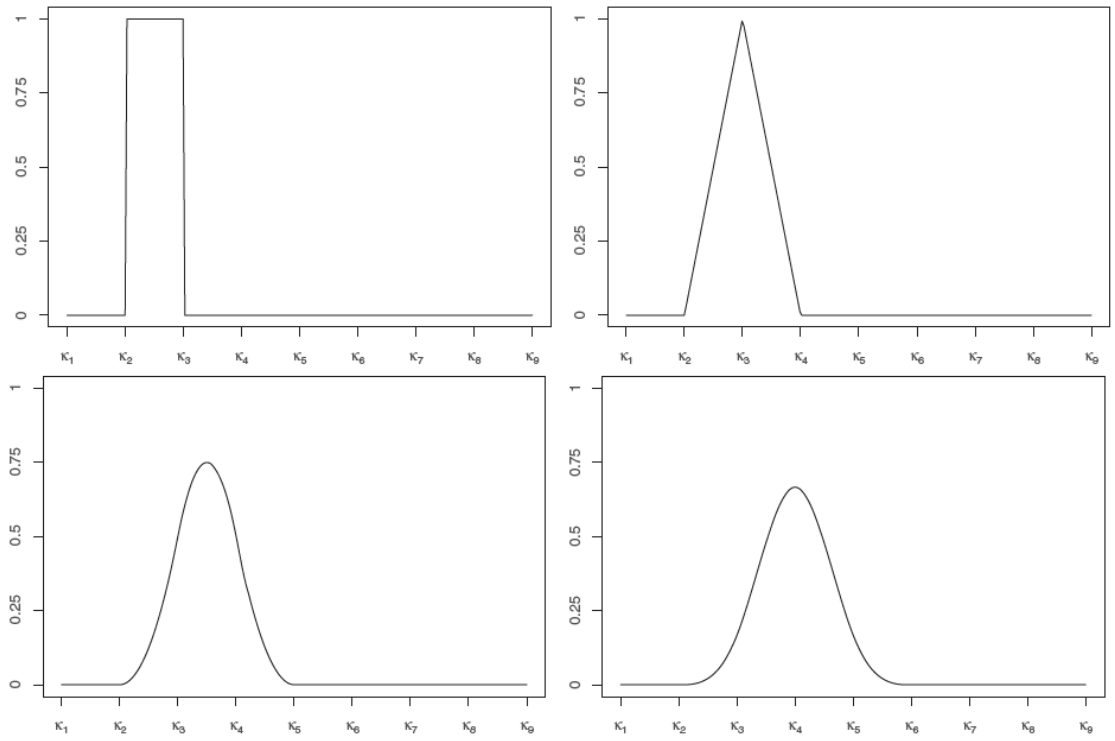

Abbildung 4: Einzelne B-Spline-Basisfunktionen vom Grad $l = 0, 1, 2, 3$ zu äquidistanten Knoten (Fahrmeir et al.; 2009, S. 303)

An jeder Stelle $z \in [a, b]$ ist die Summe der Basisfunktionen entsprechend 1. In Abbildung 5, entnommen aus (Fahrmeir et al.; 2009, S. 304), sind die vollständigen Basen der in Abbildung 4 gezeigten Basisfunktionen $l = 1, 2, 3$ zu sehen.

Die Anzahl der Knoten wird hier mit m bezeichnet. B-Spline-Basisfunktionen zum Grad $l = 1$ sind definiert durch (Fahrmeir et al.; 2009, S 305):

$$B_j^l = \frac{z - k_j}{k_{j+1} - k_j} \mathbb{1}_{[k_j, k_{j+1})}(z) + \frac{k_{j+2} - z}{k_{j+2} - k_{j+1}} \mathbb{1}_{[k_j+1, k_{j+2})}(z) \qquad j = 1, ..., m \tag{12}$$

$\mathbb{1}(z)_{[k_{j+1}, k_{j+2})}$ steht hierbei für die Indikatorfunktion, welche immer dann den Wert 1 annimmt, wenn sich das angegebene Argument z in dem angegeben Intervall $[k_{j+1}, k_{j+2})$ befindet. Die Basisfunktion ist entsprechend auf den Intervallen $[k_j, k_{j+1})$ und $[k_j + 1, k_{j+2})$ definiert. Sie wird am Knoten k_{j+1} stetig zusammengesetzt und besteht aus zwei linearen Teilstücken.
Aus diesen Basen ergibt sich $f(z)$ als Linearkombination von $d = m + l - 1$ Basis-

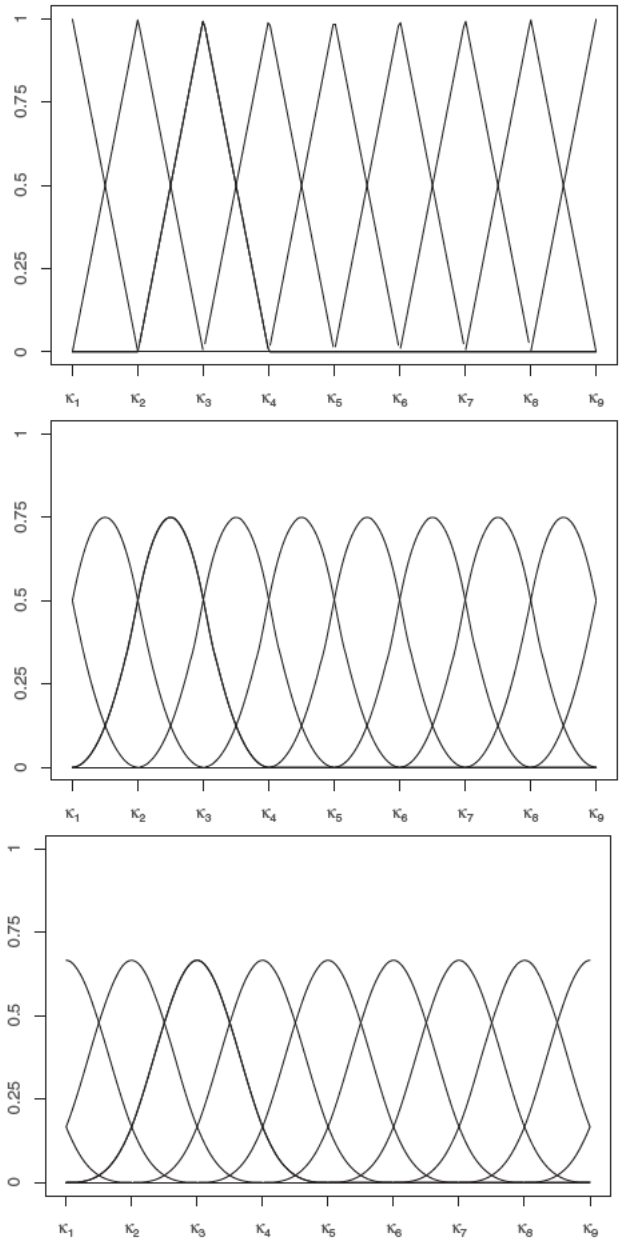

Abbildung 5: B-Spline-Basen vom Grad l = 0,1,2,3 zu äquidistanten Knoten (Fahrmeir et al.; 2009, S. 304)

funktionen mit m als Anzahl der Knoten (Fahrmeir et al.; 2009, S. 305):

$$f(z) = \sum_{j=1}^{d} \gamma_j B_j(z) \tag{13}$$

Wie bereits erwähnt hat diese Darstellung der B-Splines den Vorteil, dass es zu einem linearen Problem führt und durch eine Anpassung von Daten nach der Methode der kleinsten Quadrate (KQ) geschätzt werden kann (Blobel and Lohrmann; 2013,

S. 231). Damit die jeweilige Amplitude γ der B-Splines geschätzt werden kann, müssen zwei weitere Knoten außerhalb des Definitionsbereichs $[a, b]$ zur Knotenmenge ergänzt werden. Für die hier betrachteten äquidistanten Knoten wird der gleiche Abstand zwischen allen benachbarten Knoten genutzt. Die aus den Splines resultierende Design-Matrix $\boldsymbol{Z}$ erhält man mit (Fahrmeir et al.; 2009, S. 306):

$$\boldsymbol{Z} = \begin{pmatrix} B_1^l(z_1) & ... & B_d^l(z_1) \\ \vdots & \vdots & \vdots \\ B_1^l(z_n) & ... & B_d^l(z_n) \end{pmatrix} \tag{14}$$

Die hierbei, im Gegensatz zur Designmatrix aus 3.2.1, fehlende Konstante in der ersten Spalte erklärt sich dadurch, dass sie implizit in den übrigen Basisfunktionen erhalten ist (Fahrmeir et al.; 2009, S. 306). Durch die lokale Definition der B-Spline-Basis besteht die Designmatrix zum größten Teil aus Nullen. Die Designmatrix wird für das Schätzen des Koeffizientenvektors $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_n)$ in der Modellgleichung $\boldsymbol{y} = \boldsymbol{Z\gamma} + \boldsymbol{\epsilon}$ benötigt. Der KQ -Schätzer liefert hierbei:

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z'Z})^{-1}\boldsymbol{Z'y} \tag{15}$$

Veranschaulicht wird dies durch Abbildung 6 aus (Fahrmeir et al.; 2009, S. 307).

Anhand der Abbildung 6 lässt sich die Prozedur des Schätzens schrittweise erklären:

1. Berechnung einer vollständigen Knotenbasis zur Basis l (hier $l = 1$), siehe Abbildung 6a

2. Berechnung der Amplituden γ_j für jede Basisfunktion anhand des KQ-Schätzers, siehe Abbildung 6b

3. Berechnung der Funktion $f(z)$ durch Addition der skalierten Basisfunktionen, siehe Abbildung 6c

Bei der Wahl der Knoten muss man vorsichtig sein, da zu viele Knoten zu einer sogenannten Überanpassung führen können. Um dies zu verhindern, werden verschiedene mögliche Knotenanzahlen durch eine k-fache *Kreuzvalidierung* (Abschnitt 3.4) und den RMSE (Abschnitt 3.3) verglichen. Das Modell, welches die Daten gemäß des RMSE's am besten repräsentiert, wird gewählt. Ein Modell, welches sowohl über $\boldsymbol{\beta}$ geschätzte lineare Einflüsse, als auch über $f(\boldsymbol{z})$ geschätzte Polynomsplines verfügt,

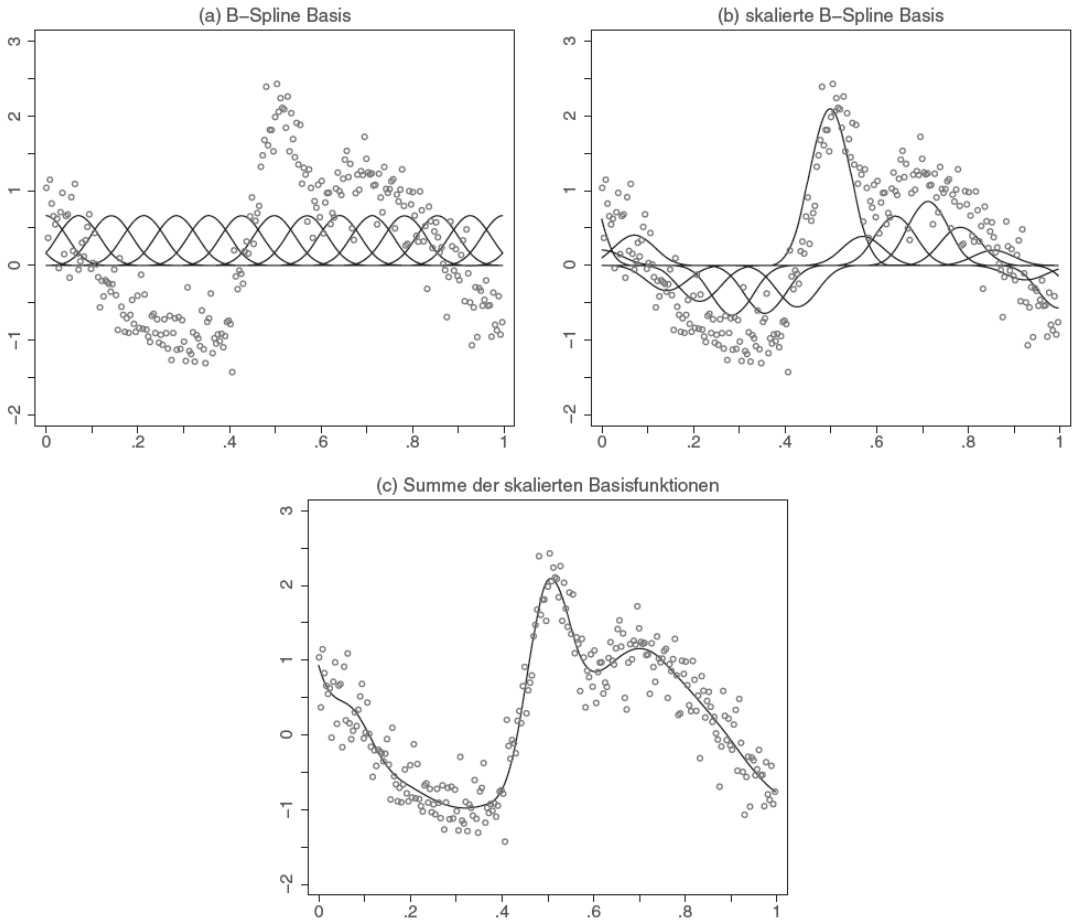

Abbildung 6: Schematische Darstellung der γ Schätzung

kombiniert diese Einflüsse und nimmt dann folgende Form an:

$$\boldsymbol{y} = \boldsymbol{X\beta} + f(\boldsymbol{z}) + \boldsymbol{\epsilon} = \boldsymbol{X\beta} + \sum_{j=1}^{d} \gamma_j B_j(z) + \boldsymbol{\epsilon} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{\epsilon} \tag{16}$$

3.2.3 Konfidenzintervalle

Bei der Schätzung $\hat{y} = E(y|x_1, ..., x_n) = \beta_0 + \beta_1 x1 + ... + \beta_k x_k$ handelt es sich um eine Punktschätzung. Für die Zielvariable y wird ein genauer Wert geschätzt. Genauso kann auch ein Intervall geschätzt werden, in welchem der Wert liegen könnte, dies nennt man eine Konfidenzschätzung bzw. Schätzung eines Konfidenzintervalls (Becker; 2005, S. 147). Ein Konfidenzschätzer liefert, basierend auf der Datengrundlage, einen Bereich $KI_\upsilon = [K_u, K_o]$ von möglichen Werten des Parameters υ. Ein solches Intervall kann man auf Basis der Daten und einer Wahrscheinlichkeit $P_\upsilon(\upsilon \in KI_\upsilon = 1 - \alpha)$ mit $\alpha \in (0, 1)$, mit welcher der Parameter υ in KI_υ liegen soll, bestimmen. Dementsprechende Konfidenzintervalle sind für die Schätzung der Zielvariable y sowie mit den β- und auch γ-Koeffizienten einer Regression möglich. Für

die jeweiligen Parameter liegen oft Verteilungsannahmen vor, welche das Erstellen einer entsprechenden Konfidenzschätzung vereinfachen. So wird beispielsweise bei der Variable X von einer Normalverteilungsannahme ausgegangen, wodurch sich das entsprechende Konfidenzintervall $P(\upsilon \in KI_\upsilon) = P(K_u \leq \upsilon \leq K_o) \geq 1 - \alpha$ ergibt (Becker; 2005, S. 152):

$$\begin{aligned} [\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}] \\ \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \end{aligned} \tag{17}$$

Dabei ist $\overline{X}$ das arithmetische Mittel und $z_{1-\alpha/2}$ das $(1-\alpha)$-Quantil der Standardnormalverteilung. Für Regressionskoeffizienten ergeben sich Konfidenzschätzungen aus der Normalverteilungsannahme (Fahrmeir et al.; 2009, S. 119):

$$[\beta_j - t_{n-p}(1-\alpha/2)se_j; \beta_j + t_{n-p}(1-\alpha/2)se_j] \tag{18}$$

Dabei ist n die Anzahl der Beobachtungen, p die Anzahl der Parameter und $t_{n-p}(1-\alpha/2)$ beschreibt das entsprechende Quantil der Student-t-Verteilung. se_j ist die geschätzte Standardabweichung.
Ein Prognoseintervall für y_0 an der Stelle $\boldsymbol{x}_0$ zum Niveau $1-\alpha$ ist gegeben durch (Fahrmeir et al.; 2009, S. 123):

$$\boldsymbol{x'}_0\hat{\boldsymbol{\beta}} \pm t_{n-p}(1-\alpha/2)\hat{\sigma}\sqrt{(1 + \boldsymbol{x'}_0(\boldsymbol{X'X})^{-1}\boldsymbol{x_0})} \tag{19}$$

3.3 Root-mean-squared error (RMSE)

Der RMSE (Root-mean-squared error) übersetzt: Die Wurzel aus dem gemittelten Fehlerquadrat dient als Maß bezüglich der Anpassungsgüte eines Modells. Dabei wird durch den RMSE ein Maß zur Modellgenauigkeit errechnet:

$$RMSE(\hat{f}(\boldsymbol{x}), \boldsymbol{x}, \boldsymbol{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{20}$$

$\hat{f}$ steht hierbei für eine Funktion (hier Regression), welche aus gegebenen Daten $\boldsymbol{x}$ versucht, die Zielvariable $\boldsymbol{y}$ zu schätzen. Der RMSE gibt dann die Wurzel aus der mittleren quadratischen Abweichung zwischen den wahren y_i, und den geschätzten $\hat{y}_i$, an. Im Rahmen dieser Bachelorarbeit wird der RMSE genutzt, um verschiedene Modelle miteinander zu vergleichen und somit das Modell zu wählen, welches besser zu den Daten passt, bei welchem entsprechend die Abweichung zwischen den

geschätzten und den vorliegenden Daten am geringsten ist. Der RMSE hat gegenüber anderen Maßen einen Vorteil bezüglich der Modellgenauigkeit, welcher darin besteht, dass er die Abweichung in derselben Skala angibt, wie jene tatsächlich gemessen wird.

3.4 k-fache Kreuzvalidierung

Um den Vorhersagefehler möglichst präzise zu schätzen, eignet sich das Verfahren der k-fachen *Kreuzvalidierung*. Dies ist notwendig, weil eine Validierung des Modells alleine anhand des RMSE aus Kapitel 3.3 dazu führen kann, dass man zwar jenes Modell wählt, welches die vorliegenden Daten am besten voraussagen würde, jedoch schlechter zu verallgemeinern ist als ein mögliches anderes Modell. Dementsprechend wird oftmals der Datensatz in einen Trainings- und Testdatensatz aufgeteilt. Der Trainingsdatensatz dient dazu, die Parameter eines Modells zu schätzen, während der Testdatensatz genutzt wird, um die Prognosegüte des Modells (hier anhand vom RMSE, Kapitel 3.3) zu schätzen. Somit wird das jeweilige, aus den Trainingsdaten geschätzte Modell, anhand der Testdaten bewertet. Dies führt dazu, dass eher ein Modell ausgewählt wird, welches sich gut verallgemeinern lässt und eine, nicht ausschließlich auf den Trainingsdaten gute Anpassung bietet.
Die k-fache *Kreuzvalidierung* verfeinert diese Überlegungen zur Modellevaluierung. Anstatt in einen Trainings- und Testdatensatz zu unterteilen, wird bei einer k-fachen *Kreuzvalidierung* jeder von k Teildatensätzen einmal als Testdatensatz genutzt, während die anderen als Trainingsdatensätze fungieren. Dabei wird der Datensatz in k möglichst gleich große Teildatensätze $\boldsymbol{X_j}$ geteilt. Jeder dieser Teile wird in einem von k Schritten einmalig als Testdatendatz genutzt, während die anderen zur Schätzung der Modellparameter genutzt werden (James; 2014, S. 241). Mittels dieser geschätzten Parameter werden also die Daten des Testdatensatzes prognostiziert und dann mittels der wahren Parameter der RMSE (siehe Kapitel 3.3) errechnet. Dementsprechend erhält man für jeden der k Teildatensätze einen Wert $RMSE_j$ für den RMSE des anhand der anderen Datensätze geschätzten Modells. Aus diesen k Werten für den RMSE bildet man also den Mittelwert, um den mittleren RMSE zu erhalten:

$$\overline{RMSE} = \frac{1}{k}\sum_{j=1}^{k} RMSE_j = \frac{1}{k}\sum_{j=1}^{k}\sqrt{\frac{1}{n_j}\sum_{i=1}^{n_j}(y_{i,j} - \hat{y}_{i,j})^2} \tag{21}$$

Dieses Verfahren kann optimal genutzt werden, um so genannte Tuning-Parameter (hier τ) in einem möglichen Modell zu optimieren. Für jeden möglichen Tuning-Parameter wird eine *Kreuzvalidierung* durchgeführt. Der Tuningparameter, bei wel-

chem der mittlere RMSE ($\overline{RMSE}$) am niedrigsten ist, wird dann als optimaler Tuningparameter gewählt.
Dieser Algorithmus ist hier schrittweise zusammengefasst:

Tabelle 3: k-fache *Kreuzvalidierung* für Parameter τ

Teile den Datensatz in k gleich große Teildatensätze $\boldsymbol{X}_j$
Für jedes τ in $[\tau_{min}, ..., \tau_{max}]$
Für jedes j in $[1, ..., k]$
Verwende Teildatensatz $\boldsymbol{X}_j$ als Testdatensatz
Verwende die restlichen $k-1$ Teildatensätze $\boldsymbol{X}_{1,...,j-1,j+1,...,k}$ und τ, um die Regressionsparameter zu schätzen
Prognostiziere $\hat{y}_{test}$ des Testdatensatzes $\boldsymbol{X}_j$ anhand der Parameterschätzungen
Berechne den $RMSE_j$ anhand der wahren y-Werte
Erreichen des letzten j
Bilde ($\overline{RMSE}_\tau$) der k $\boldsymbol{X}_j$
Erreichen des letzten τ
Bestimme das optimale τ: $\tau_{opt} = argmin_\tau \overline{RMSE}_\tau$

3.5 Bootstrap

Beim *Bootstrap*-Verfahren handelt es sich um eine Resampling-Methode. Das *Bootstrap*-Verfahren ist eine statistische Technik, um unbekannte Verteilungen abzuschätzen (Boos; 2013, S. 413). Der Vorteil der *Bootstrap*-Methode ist, dass es sich um ein non-parametrisches Verfahren handelt, welches keine Verteilungsannahme trifft und kann dadurch auch angewandt werden, wenn Normalverteilungsannahmen des Modells nicht oder nur zweifelhaft erfüllt sind. Dies wird oft bei der Ermittlung von Standardabweichungen der Fehler ϵ_i oder bei der Abschätzung von Prognoseintervallen angewandt (Boos; 2013, S. 413). Beim *Bootstrap*-Verfahren werden n_{sim}-viele *Bootstrap*-Zufallsstichproben $\boldsymbol{X}_i^*$ der Größe $n = |\boldsymbol{X}|$ mit Zurücklegen gezogen. $|\boldsymbol{X}|$ ist hierbei die Kardinalität von $\boldsymbol{X}$, also die Anzahl der vorliegenden Beobachtungen. Die Anzahl n_{sim} liegt typischerweise zwischen 500 und 2000 (Bhattacharya; 2016, S. 258). Für jede *Bootstrap*-Zufallsstichprobe $X_i^* \in [\boldsymbol{X}_1^*, ..., \boldsymbol{X}_{nsim}^*]$ wird dann die interessierende Statistik $T_i(\boldsymbol{X}_i^*) = T_i(x_{i,1}^*, ..., x_{i,n}^*)$ (z.B.: die Standardabweichung der Residuen oder Regressionskoeffizienten) berechnet. Somit ergeben sich n_{sim}-viele Schätzungen der interessierenden Statistik $T_i(\boldsymbol{X}_i^*)$ und damit eine empirische Verteilung. Aus dieser Verteilung kann man für jeden der geschätzten Parameter den Standardfehler schätzen und aus den $(\alpha/2)$-Quantilen ein entsprechendes Konfidenzintervall bestimmen (Bhattacharya; 2016, S. 258).

3.6 Inverse Schätzung

3.6.1 Inverse Punktschätzung

Das Ziel beim inversen Schätzen ist es, für eine vorgegebene Zielvariable Y_0 eine Einflussvariable X_0 vorherzusagen. In der Statistik bezeichnet man dies auch als Kalibrierungsproblem oder inverse Regression (Draper and Smith; 2014, S. 83). In diesem Kapitel wird erläutert, auf welcher mathematischen Grundlage Rückschlüsse von der Ziel- auf die Einflussvariable getroffen werden können und wie sich die daraus resultierende Schätzgenauigkeit ergibt.
Im Folgenden wird ein bereits geschätztes Regressionsmodell $\hat{y}_i = f(x_i, \hat{\boldsymbol{\Theta}})$ betrachtet. $\hat{\boldsymbol{\Theta}}$ bezeichnet dabei den Parameterraum, welcher beispielsweise geschätzte $\hat{\boldsymbol{\beta}}$ oder im Falle von B-Splines $\hat{\boldsymbol{\gamma}}$ enthält. $f(x_i, \hat{\boldsymbol{\Theta}})$ ist dabei eine über das betrachtete X_0-Intervall monoton (Greenwell and Kabban; 2014, S. 90). Die Schätzung wird in Abbildung 7 veranschaulicht.

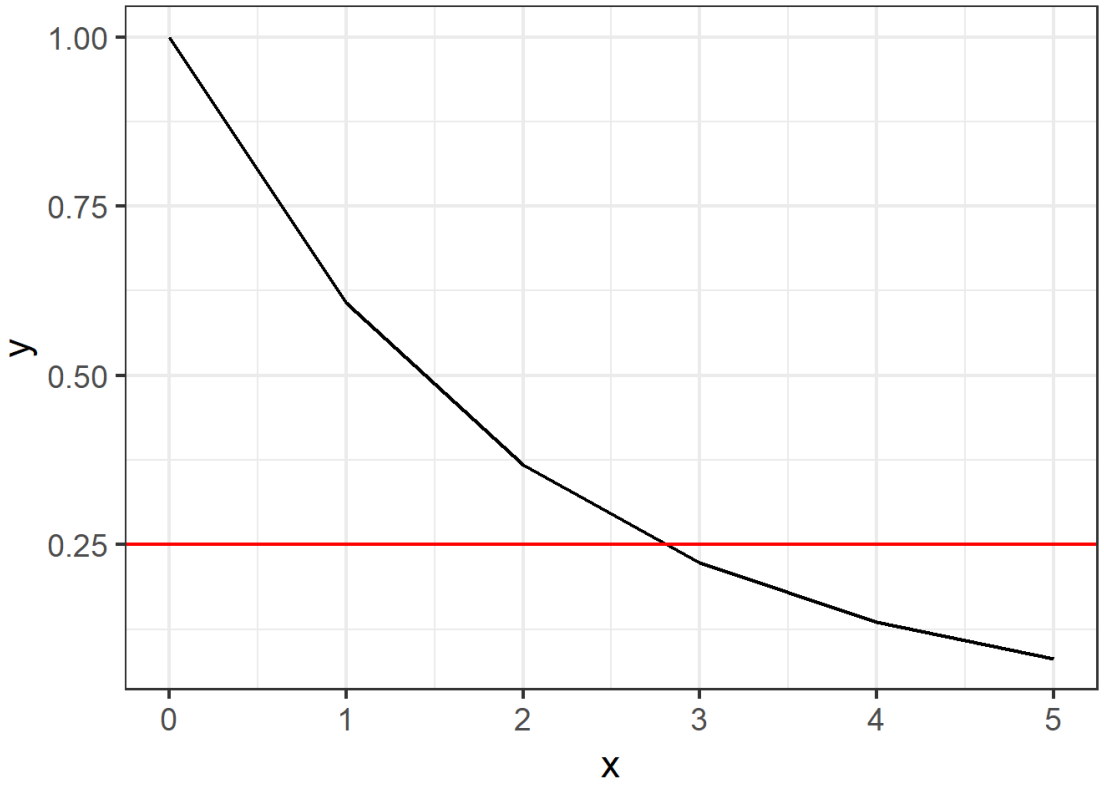

Abbildung 7: Es wird eine Schätzung von x auf y betrachtet. Beim Invertieren wird beispielsweise ein Wert für $Y_0 = 0.25$ gesucht. Die Punktschätzung von X_0 für $Y_0 = 0.25$ entspricht also dem Wert, bei dem die geschätzte Regressionsgerade den Wert $Y_0 = 0.25$, hier durch die rote Linie gekennzeichnet, trifft.

Eine Punktschätzung für X_0 erhält man mathematisch durch das Lösen der Gleichung $Y_0 = f(x_i, \hat{\boldsymbol{\Theta}}) + \epsilon_i$ nach X_0. Dies wird im Folgenden immer auf diese Art und Weise berechnet:

$$\begin{aligned} Y_0 = f(x_i, \hat{\boldsymbol{\Theta}}) + \epsilon_i \Leftrightarrow Y_0 - f(x_i, \hat{\boldsymbol{\Theta}}) = \epsilon_i \Leftrightarrow \\ \text{da der Erwartungswert } E(\epsilon_i) = 0 \\ \Leftrightarrow Y_0 - f(x_i, \hat{\boldsymbol{\Theta}}) = 0 \\ \text{mit } x_i = \begin{pmatrix} x_{i,1} & ... & X_{i,0} & ... & x_{i,k} \end{pmatrix}' \end{aligned} \tag{22}$$

Der Datenvektor $x_{i,j} = \begin{pmatrix} x_{i,1} & ... & X_{i,0} & ... & x_{i,k} \end{pmatrix}'$ besitzt dabei für jedes $j \in [1, ..., k]$ außer $X_{i,0}$ einen Wert. $X_{i,0}$ bezeichnet hierbei den Wert, welcher durch die Regressionsgleichung und dem gegebenen Y_0 bestimmt werden soll. Eben jener wird in dieser Abschlussarbeit anhand der genannten Formel für alle $X_0 \in [X_{0,min}, X_{0,max}]$ berechnet. Ein entsprechender Wert, welcher diese Gleichung löst, wird als $\hat{X}_0$ gewählt. Die Berechnung von Punktschätzern wird noch durch die Berechnung von Prognoseintervallen ergänzt, welche die aus der Berechnung resultierende Unsicherheit quantifizieren.

3.6.2 Inverse Prognoseintervalle

Die inversen Prognoseintervalle werden hierbei mittels *Bootstrap* (Kapitel 3.5), einer Alternative zu herkömmlichen Schätzmethoden, ermittelt. Dabei werden gemäß Abschnitt 3.5 aus der Verteilung der Daten $\boldsymbol{X}$, mittels welcher das Regressionsmodell geschätzt wurde, n_{sim}-viele Zufallsstichproben gezogen. Die Stichprobengröße entspricht dabei $|X|$ der Anzahl der Beobachtungen von X. Die Stichprobe unterscheidet sich dadurch von X, dass sie mit zurücklegen gezogen wurde und dementsprechend Beobachtungen mehrfach in die Stichprobe gelangen können, während andere möglicherweise nicht gezogen werden. Für jede der n_{sim}-vielen Stichproben wird dann für ein gegebenes Y_0 ein X_0 durch eine Punktschätzung geschätzt. Die geschätzten X_0 bilden wiederum eine Verteilung. Aus dieser Verteilung wird eine Standardabweichung berechnet:

$$\begin{aligned} SD_{X_0} = |\sqrt{\frac{1}{n_{sim} - 1} \sum_{i=1}^{n_{sim}} (\hat{X}_i - \overline{X})^2}|, \\ \text{mit } |\cdot| \text{ als Betrag} \end{aligned} \tag{23}$$

Anhand der empirischen Verteilung von X_0 wird also die „wahre“ Verteilung von X_0 approximiert. Anhand dieser Approximation werden entsprechende Prognoseintervalle berechnet. Approximativ ergeben sich jene bei beliebiger Verteilung und geschätzter Standardabweichung zu (Becker; 2005, S. 159):

$$KI_v = [\overline{X} \pm t^*_{n-1;(1-\alpha/2)} \frac{\hat{SD}}{\sqrt{n}}] \tag{24}$$

In diesem Fall ist $\overline{X} = \overline{X_0}$ und $n = n_{sim}$. Da im Rahmen dieser Abschlussarbeit stets $n_{sim} > 30$ gewählt wird, können, statt der mit dem * gekennzeichneten $t^*_{n-1;(1-\alpha/2)}$ Quantile der t-Verteilung, die Quantile $z_{(1-\alpha/2)}$ der Standardnormalverteilung verwendet werden. Im späteren Verlauf dieser Abschlussarbeit wird ein Wert gesucht, für den gilt $F_{0.95}(X_0) : F_{0.95}(X_0) = P(X_0 \geq \tilde{X_0}) \geq 0.95$ mit $F(\cdot)$ als Verteilungsfunktion und $\tilde{X_0} = 0$ als „wahren“ Wert für X_0. Dies entspricht der oberen Grenze des einseitigen $1 - \alpha = 0.95$-Konfidenzintervalls $KI_v = (-\infty; \overline{X} + z_{1-\alpha} \cdot \frac{SD}{\sqrt{n}})$ mit $z_{1-\alpha} = 1.65$ für $\alpha = 0.05$. Aus diesen mathematischen Grundlagen lässt sich nun eine Schätzung für X_0 konstruieren, welche zu 95% $Y_0 \geq f(x_i, \hat{\mathbf{\Theta}})$ mit $X_0 \in x_i$ erfüllt.

4 Auswertung

Aufbauend auf der in Kapitel 3 erläuterten methodischen Grundlage wird nun die durchgeführte Parameterstudie dargelegt. Der zugrundeliegende Versuchsplan und die dahinterstehende Theorie wurden bereits in Abschnitt 2.3 bzw. Abschnitt 3.1 erläutert. Im ersten Schritt erfolgt eine Auseinandersetzung mit der Rechenzeit und ihrer Abhängigkeit von den numerischen Parametern. Aufbauend auf den dabei gewonnenen Erkenntnissen wird visualisiert, wie die physikalischen und numerischen Parameter auf die Zielvariable „relative Abweichung der numerischen zur analytischen Lösung bezüglich des Porendrucks“ ($error_p$) wirken. Dies wird durch entsprechende Variablentransformationen unterstützt. Darauf aufbauend wird dies modelliert, indem die verschiedenen Parametereinflüsse passend mit einbezogen werden. Das daraus resultierende Modell wird anhand ausgewählter Qualitätsmerkmale bewertet und anschließend verwendet, um den Rückschluss von der Zielvariable $error_p$ auf die numerischen Parameter dwn_{min} und nz zu errechnen. Diese inverse Regression wird schließlich anhand ihrer Schätzgenauigkeit bewertet. Dafür werden entsprechende Schätzungen der numerischen Parameter, abhängig von physikalischen Parametern, in Bezug auf ihre Genauigkeit betrachtet.

4.1 Datensituation

Im Rahmen dieser Abschlussarbeit wurden weit über 500.000 Datenpunkte erzeugt und betrachtet. Auf Grund der gewonnenen Informationen über die Daten, mussten fortlaufend Datenpunkte mit neuen Anforderungen erzeugt werden. Beispielsweise das in Abschnitt 4.2 erläuterte Minimalsetzen des Parameters „minimaler Wert der Länge der Diffusion Dwn_{min}“ führte zu neuen Anforderungen an die Variablen. Für die abschließende Modellierung und Visualisierung der Problemstellung wurden 320.000 Datenpunkte erzeugt. Davon wurden 80%, also 256.000 Datenpunkte, als Trainingsdatensatz und die anderen 20%, somit 64.000 Datenpunkte, als Testdatensatz genutzt. Für die Berechnung des linearen Modells bei den *Bootstrap*-Simulationen wurde auf Grund der hohen Rechenzeit ein Modell, welches aus 100.000 Datenpunkten geschätzt wurde verwendet.

4.2 Rechenzeit

Die Anforderung an die numerischen Parameter ist, bei möglichst geringer Rechenzeit, die maximale Fehlerschranke von 1.00E−06 nicht zu überschreiten. Dafür wird im ersten Schritt ein Modell betrachtet, welches den Einfluss, den die numerischen Parameter auf die Rechenzeit haben, modelliert. Da die Berechnungen auf verschiedenen Computern mit einer unterschiedlichen Leistung und Anzahl an Kernen getätigt wurden, wird sich bei der Betrachtung der Rechenzeit ausschließlich auf die Berechnungen durch einen Server des Leibniz-Rechenzentrums in München konzentriert. Die anderen Berechnungen waren in Bezug auf ihre Rechenzeit durch schwankende Leistungen, ausgelöst durch paralleles Nutzen der Computer, verzerrt.
Auffällig hierbei ist, dass zwischen der numerischen Variable „Anzahl an Gridpunkten“ (nz) und der Berechnungsdauer ein nahezu perfekter linearer Zusammenhang vorliegt. Dies untermauert auch die Visualisierung dessen in Abbildung 8.

Der minimale Wert der Länge der Diffusion (Dwn_{min}) hingegen, scheint keinen Einfluss auf die Berechnungsdauer zu haben. Dies wird zum einen aus der Visualisierung in Abbildung 8 deutlich, zum anderen wird dies in einem Regressionsmodell, welches die Rechenzeit anhand der Anzahl an Gridpunkten sowie dem minimalen Wert der Länge der Diffusion modelliert, deutlich. Anhand dessen wurde in Absprache mit dem Projektpartner aus dem Department für Geo- und Umweltwissenschaften (Geophysik) der Ludwig-Maximilians-Universität entschieden, dass der numerische Parameter „minimaler Wert der Länge der Diffusion“ (Dwn_{min}) auf seinen Minimalwert ($\frac{1.00\text{E}{-16}}{hwid}$) gesetzt wird. Der zukünftig angenommene minima-

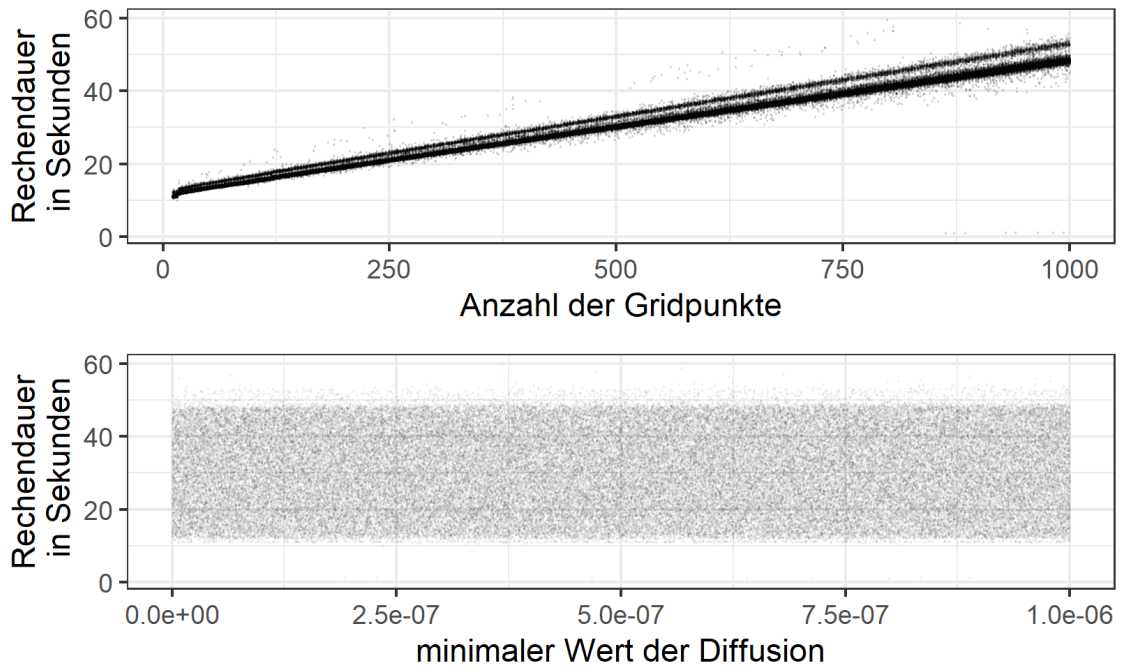

Abbildung 8: Rechenzeit nach der Anzahl an Gridpunkten nz und dem minimalen Wert der Länge der Diffusion Dwn_{min}

le Wert von Dwn_{min} entspricht dem Wert, welcher die Zielvariable $error_p$ minimiert. Dadurch hängt Dwn_{min} deterministisch von dem physikalischen Parameter „Breite der Bruchzone“ ($hwid$) ab. Aus dieser Abhängigkeit folgt, dass zwei weitere numerische Variablen deterministisch vom physikalischen Parameter „Breite der Bruchzone“ ($hwid$) abhängen. Der „maximale Wert der Länge der Diffusion“ ($Dwn_{max} = \frac{10}{hwid}$) hängt direkt von der Breite der Bruchzone ($hwid$) ab. Der Rasterabstand ($dlDwn = \frac{ln(Dwn_{min}/Dwn_{max})}{nz-1}$) hängt von Dwn_{min} und Dwn_{max} ab, welche wiederum deterministisch von der Breite der Bruchzone ($hwid$) abhängen. Folglich bleibt der einzige, bezüglich der Rechenzeit zu optimierende numerische Parameter „die Anzahl an Gridpunkten“ (nz). Zusätzlich spricht für dieses Vorgehen, dass die Berechnung der Rechenzeit, vor und nach dieser Minimalsetzung von dem minimalen Wert der Länge der Diffusion Dwn_{min}, identisch blieb. Die Rechenzeit des Servers ergibt sich anhand einer linearen Einfachregression zu:

$$\widehat{Rechenzeit}_{Sekunden} = 12_{Sekunden} + nz \cdot 0.03724_{Sekunden} \tag{25}$$

4.3 Modellierung

4.3.1 Modellierung des relativen Fehlers

Die Modellierung der relativen Abweichung zwischen numerischer und analytischer Lösung bezüglich des Porendrucks erfolgt über ein Regressionsmodell. Bei diesem Modell wird versucht, die Einflüsse der physikalischen Parameter α_{th}, α_{hy}, λ und *hwid*, sowie des numerischen Parameters nz in Bezug auf die Vorhersage der Einflussvariable $error_p$ optimal zu modellieren. Um den Einfluss der Variablen optimal zu bestimmen, bietet es sich an, diese zu visualisieren. Eine solche Visualisierung befindet sich in Abbildung 9.

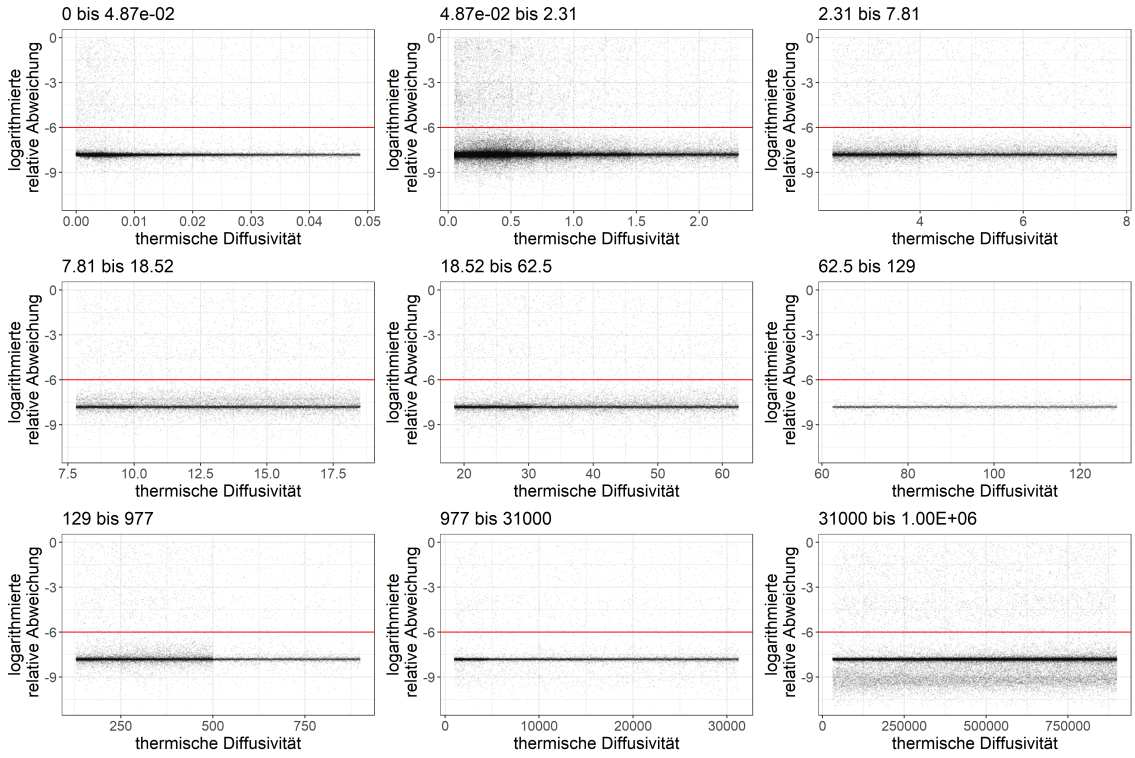

Abbildung 9: Einfluss der thermischen Diffusivität auf die relative Abweichung zwischen numerischer und analytischer Lösung bezüglich des Porendrucks

Bei der Betrachtung des Einflusses der thermischen Diffusivität auf die relative Abweichung zwischen numerischer und analytischer Lösung bezüglich des Porendrucks (siehe Abbildung 9) fällt auf, dass der Einfluss linear und nahezu konstant ist. Die hierbei eingezeichnete rote Linie ist die gegebene Fehlerschranke (1.00E−06). Die relative Abweichung zwischen numerischer und analytischer Lösung bezüglich des Porendrucks ist logarithmiert zur Basis 10 dargestellt. Der Plot wurde in verschiedene Achsenabschnitte eingeteilt, da, wie auch hierbei ersichtlich, verschieden viele Punk-

te in den jeweiligen Intervallen mit dem genutzten *Latin-Hypercube-Sample* (Kapitel 2.3) erzeugt wurden. Die Einflüsse der physikalischen Parameter der hydraulischen Diffusivität und der Druckänderung pro Temperaturanstieg sehen ähnlich aus und eine entsprechende Visualisierung befindet sich im Anhang dieser Abschlussarbeit (siehe Kapitel 5). Um die optimale Modellierung des Einflusses auf die relative Abweichung zwischen numerischer und analytischer Lösung bezüglich des Porendrucks zu modellieren, wurde durch eine k-fache *Kreuzvalidierung* mit $k = 10$ (siehe Abschnitt 3.4) jede Kombination an möglichen Modellierungen durch Logarithmieren der Variablen miteinander anhand der mittleren Wurzel der mittleren quadratischen Abweichung verglichen. Dies ergab, dass α_{th}, die thermische Diffusivität, α_{hy}, die hydraulische Diffusivität sowie λ, die Druckänderung pro Temperaturanstieg logarithmiert modelliert werden.
Die Einflüsse der Breite der Bruchzone und der Anzahl an Gridpunkten sieht man in der folgenden Visualisierung 10:

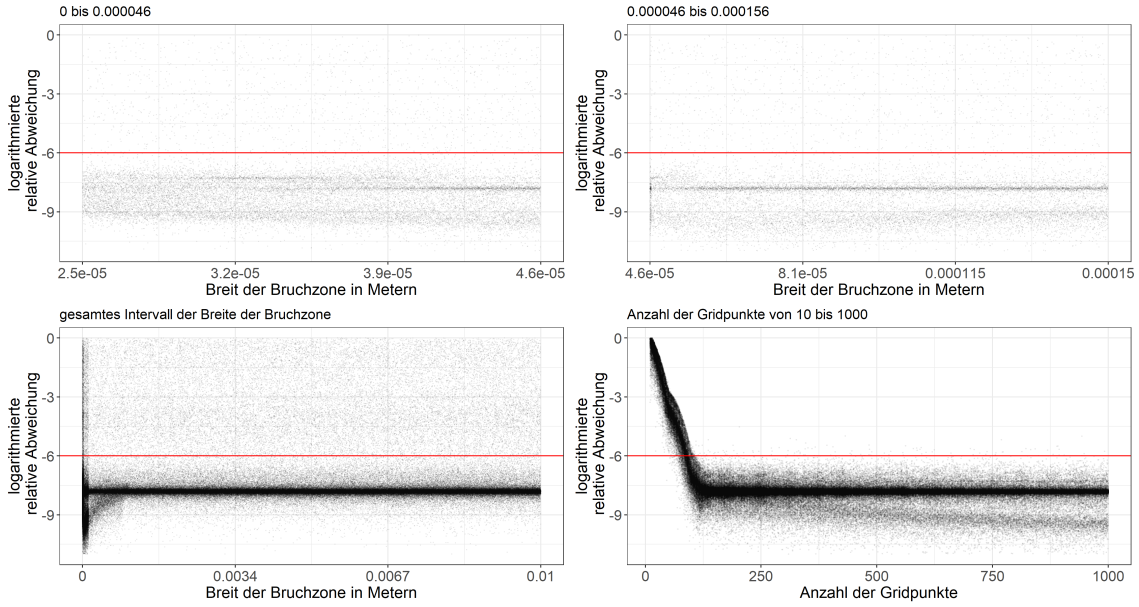

Abbildung 10: Einfluss der Bruchzonenbreite und der Anzahl an Gridpunkten auf die relative Abweichung zwischen numerischer und analytischer Lösung bezüglich des Porendrucks.

Man sieht, dass der Einfluss hier nicht auf dem ganzen Intervall linear ist. Das führt zur Anwendung der in Kapitel 3.2.2 eingeführten B-Splines. Jene werden genutzt, um den nichtlinearen Zusammenhang auf ein lineares Problem zurückzuführen. Dabei wurde die Anzahl an Knoten m und der Grad l durch eine k-fache *Kreuzvalidierung* (siehe Kapitel 3.4) bestimmt. Im Folgenden wird immer vereinfacht von einer *Kreuzvalidierung* geschrieben, damit wird die „k-fache *Kreuzvalidierung*“ mit $k = 10$

impliziert. Es wurden jeweils die Anzahl an Knoten m sowie der Grad l gewählt, bei welchem jeweils die in der durchgeführten *Kreuzvalidierung* mittlere Wurzel der mittleren quadratischen Abweichung (RMSE, siehe Kapitel 3.3) am niedrigsten war. Diese B-Spline Schätzung ist in Grafik 11 dargestellt.

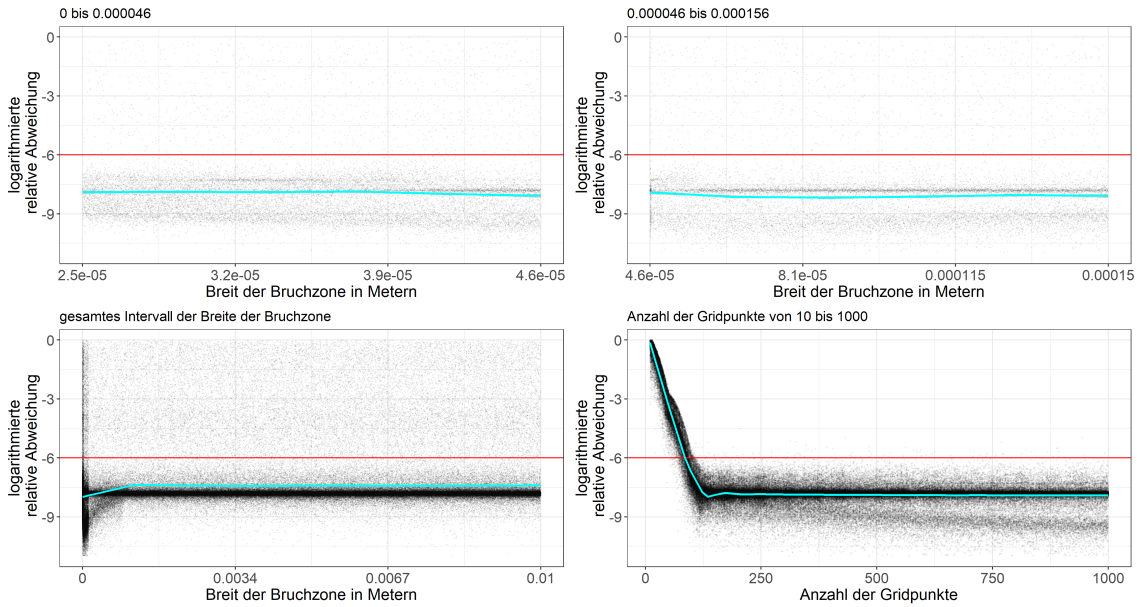

Abbildung 11: B-Splineschätzungen für die Breite der Bruchzone und die Anzahl an Gridpunkten

Die B-Spline Schätzungen für die Breite der Bruchzone *hwid* und die Anzahl an Gridpunkten nz sind in Türkis dargestellt. Es ergab sich, dass der RMSE für die Breite der Bruchzone am niedrigsten war, wenn die äquidistanten Knoten den Grad $l = 1$ annehmen und die Anzahl an Knoten mit $m = 4$ gewählt wird. Für die Anzahl an Gridpunkten ergaben sich über die angesprochene *Kreuzvalidierung* der Grad $l = 1$ und $m = 24$ Knoten. Diese Splineschätzungen wurden im Modell zur Fehlermodellierung verwendet. Die verschiedenen Modellierungsmöglichkeiten der Variablen, wie beispielsweise das Logarithmieren oder die Verwendung von Splines, wurden durch *Kreuzvalidieren* auf mögliche Interaktionen überprüft. Dabei wurde jeweils das Modell mit Interaktionen mit dem ohne Interaktionen verglichen und anhand des aus der *Kreuzvalidierung* resultierenden Ergebnisses des RMSE's entschieden, ob eine Interaktion in das Modell aufgenommen werden sollte. Dabei ergab sich, dass jede mögliche Interaktion den RMSE des Modells verbesserte. Durch die Interaktionen stieg der Rechenaufwand des Modells enorm. Folglich ergab sich aus diesen Schritte das vollständige Modell. Dieses Modell kombiniert also die zur Ba-

sis 10 logarithmierten Einflüsse von α_{th}, α_{hy} und λ mit dem über B-Splines mit 4 Basisfunktionen zum Grad $l = 1$ dargestellten Einfluss von *hwid* sowie dem über B-Splines mit 24 Basisfunktionen zum Grad $l = 1$ dargestellten Einfluss von nz additiv:

$$\begin{aligned} log_{10}(error_p) = Intercept + \beta_{\alpha_{th}} \cdot log_{10}(\alpha_{th}) + \beta_{\alpha_{hy}} \cdot log_{10}(\alpha_{hy}) \\ +\beta_\lambda \cdot log_{10}(\lambda) + \boldsymbol{\gamma}_{hwid} \cdot \boldsymbol{Z}_{hwid} + \boldsymbol{\gamma}_{nz} \cdot \boldsymbol{Z}_{nz} + \epsilon \\ \text{mit } \boldsymbol{Z}_{nz} = \begin{pmatrix} B_1(z_1) & ... & B_{24}(z_1) \\ \vdots & \ddots & \vdots \\ B_1(z_n) & ... & B_{24}(z_n) \end{pmatrix} \text{ und } \boldsymbol{\gamma}_{nz} = (\gamma_1, ..., \gamma_{24})' \\ \text{wobei } (z_1, ..., z_n) \text{ hier die Ausprägungen des Parameters } nz \text{ sind,} \\ \boldsymbol{Z}_{hwid} \text{ analog} \end{aligned} \tag{26}$$

Zusätzlich zu der hier formulierten Modellgleichung wurden noch alle Interaktionen aufgenommen. Jene fehlen hier, da insbesondere durch die Spline-Terme das Ausformulieren einer solchen Modellgleichung an Übersichtlichkeit einbüßt und nicht zielführend sein würde. Insbesondere fällt in der Abbildung 10 auf, dass beim Einfluss der Anzahl an Gridpunkten im unteren Bereich eine zweite lineare Tendenz zu sehen ist. Diese Tendenz ist auf den Einfluss von niedrigen Bruchzonenbreite-Werten zurückzuführen, welche für einen niedrigeren relativen Fehler zwischen analytischer und numerischer Lösung sprechen. Dieser Zusammenhang ist gesondert durch eine Farbunterteilung der Punkte, je nach Wert der Bruchzonenbreite, in Abbildung 12 visualisiert.

4.3.2 Modelldiagnostik

Das Modell wird nun anhand von verschiedenen Gesichtspunkten untersucht, dabei wird insbesondere die Residuenstruktur betrachtet. Dafür werden die geschätzten Werte $\widehat{error}_{p_i}$ des Modells und die dazugehörigen Residuen ϵ_i visualisiert. Da die Residuen in einem linearen Regressionsmodell unabhängig sein sollen, ist dabei im Optimalfall keine Struktur erkennbar. Außerdem sollte die Streuung für die jeweils geschätzten Werte $\widehat{error}_{p_i}$ für alle ϵ_i gleich sein, sonst scheint die Varianzhomogenität verletzt zu sein.

Zum einen fällt bei der Betrachtung der Residualstruktur in Abbildung 13 auf, dass eine lineare Struktur diagonal durch den Datenmittelpunkt geht. Dies weißt auf eine Verletzung der Modellannahmen hin. Die Residuen müssen laut jener unabhängig

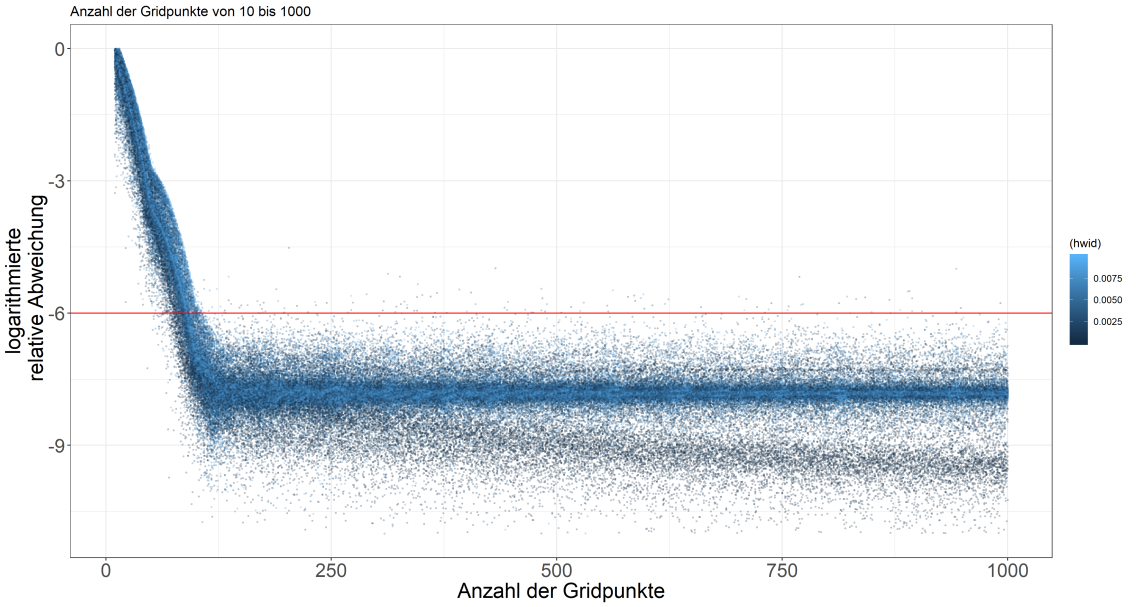

Abbildung 12: Einfluss der Anzahl der Gridpunkte und Bruchzonenbreite. Je dunkler die Punkte, desto niedriger ist der Wert für die Bruchzonenbreite (*hwid*). Scheinbar ist auch die niedrigere relative Abweichung zwischen numerischer und analytischer Lösung bezüglich des Porendrucks dadurch zu erklären.

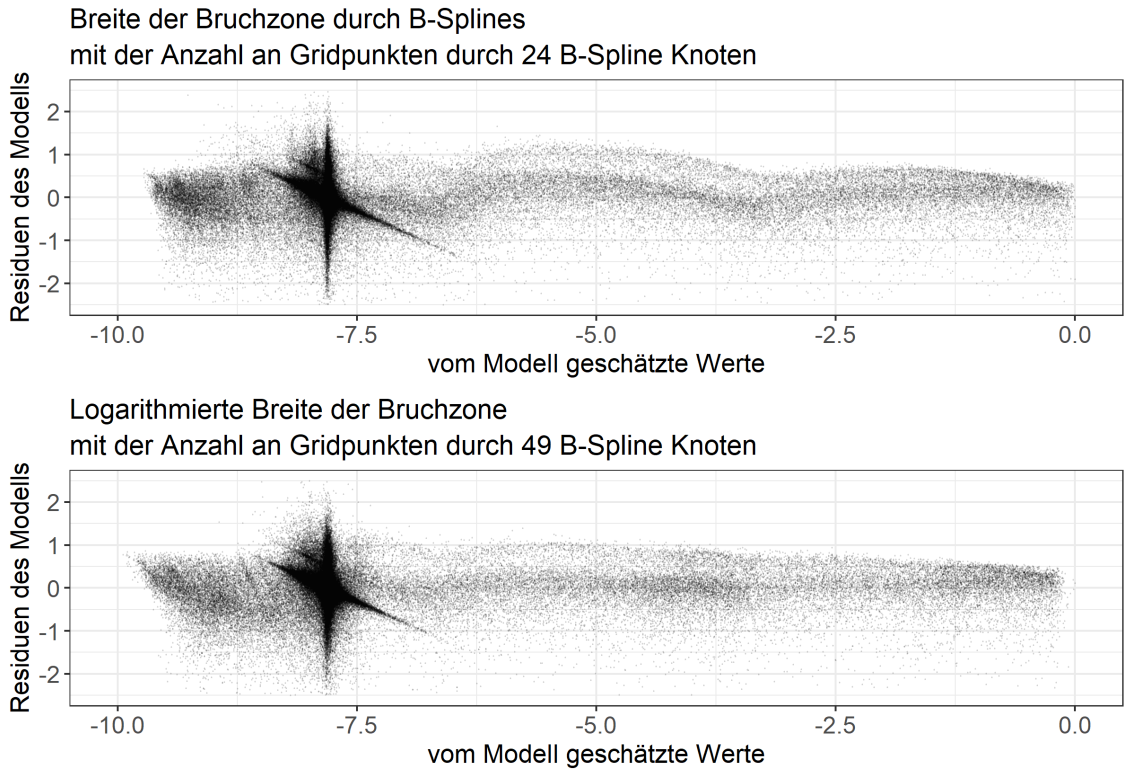

Abbildung 13: Vom Modell geschätzte Werte und die zugehörigen Residuen

sein. Im unteren Plot der Abbildung 13 findet man den Residualplot, wenn die Breite der Bruchzone (*hwid*) logarithmisch aufgenommen wird und nz über B-Splines mit $m = 49$ Knoten zum Grad $l = 1$ aufgenommen wird. Es fällt auf, dass unter den Modellspezifikationen des Modells auf der unteren Seite die Residuenstruktur ab einem Wert von -7.5 abnimmt. Im Folgenden wird auf Grund dessen jenes Modell genutzt, da es zu weniger systematischen Fehlern führt. Ein weiterer Vorteil ist, dass Interaktionen zwischen verschiedenen, über B-Splines modellierten Einflüssen sehr rechenintensiv sind, da die Interaktionen zwischen den einzelnen Basisfunktionen betrachtet werden. Das daraus resultierende Modell ergibt sich also zu:

$$\begin{aligned} log_{10}(error_p) = Intercept + \beta_{\alpha_{th}} \cdot log_{10}(\alpha_{th}) + \beta_{\alpha_{hy}} \cdot log_{10}(\alpha_{hy}) \\ +\beta_{\lambda} \cdot log_{10}(\lambda) + \beta_{hwid} \cdot log_{10}(hwid) + \boldsymbol{\gamma}_{nz} \cdot \boldsymbol{Z}_{nz} + \epsilon \\ \text{mit } \boldsymbol{Z}_{nz} = \begin{pmatrix} B_1(z_1) & ... & B_{29}(z_1) \\ \vdots & \ddots & \vdots \\ B_1(z_n) & ... & B_{49}(z_n) \end{pmatrix} \text{ und } \boldsymbol{\gamma}_{nz} = (\gamma_1, ..., \gamma_{49})' \end{aligned} \tag{27}$$

wobei $(z_1, ..., z_n)$ hier die Ausprägungen des Parameters nz sind

Um nun die Datenpunkte auf der Diagonalen durch den Datenmittelpunkt, welche systematisch über- bzw. unterschätzt werden, zu extrahieren werden jene nun rot markiert. Dieselben Datenpunkte werden jetzt in den Streudiagrammen zwischen den verschiedenen numerischen und physikalischen Parametern und dem relativen Fehler zwischen numerischer und analytischer Lösung betrachtet. In Abbilung 14 wird der Einfluss der Breite der Bruchzone und der rot markierten Daten, mit den systematischen Abweichungen, betrachtet. Die zugehörigen Visualisierungen der anderen Parameter zeigten weniger, teilweise auch keine, Auffälligkeiten und befinden sich im Anhang (siehe Abschnitt 5) dieser Abschlussarbeit.

Hierbei fällt auf, dass die Datenpunkte, welche zu einer systematischen Über- bzw. Unterschätzung führen, außerhalb des Bereichs liegen, in welchem die Breite der Bruchzone $hwid \approx 0.05$ annimmt. Bei der Betrachtung des zugehörigen Plots der partiellen Residuen in Abbildung 15 fällt zusätzlich auf, dass die Datenanpassung des Regressionsmodells beim Einfluss der Breite der Bruchzone im Bereich der $log_{hwid} \approx -4.5$ fehlspezifiziert scheint. Dies wiederum könnte durch die bereits angesprochenen B-Splines für die Bruchzonenbreite gelöst werden. Jedoch scheint das Problem in Abbildung 13, welche eine Modellierung von der Bruchzonenbreite über B-Spline beinhaltet, auch zu bestehen. Die partiellen Residuen werden hier ausschließlich für

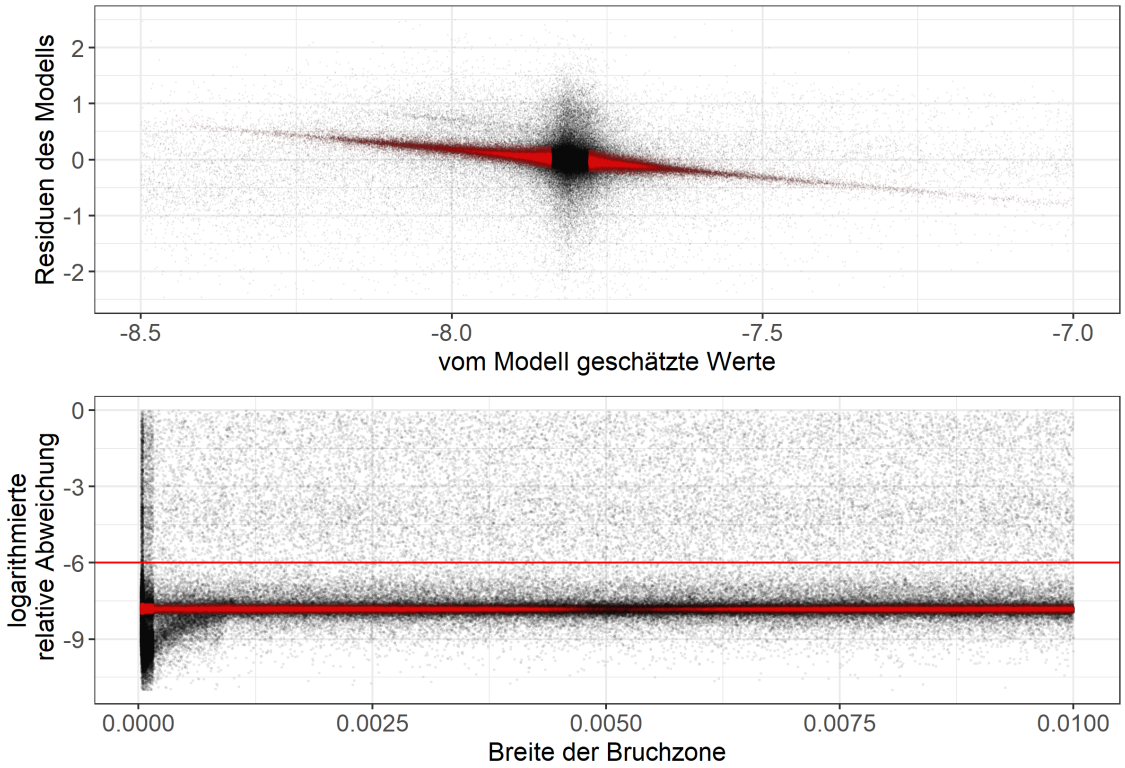

Abbildung 14: Einfluss von der Bruchzonenbreite auf den relativen Fehler zwischen numerischer und analytischen Lösung. Die Verteilung der Datenpunkte, welche zu einer systematische Über- bzw. Unterschätzung führen, sind hier rot markiert.

den Haupteffekt der Breite der Bruchzone betrachtet.

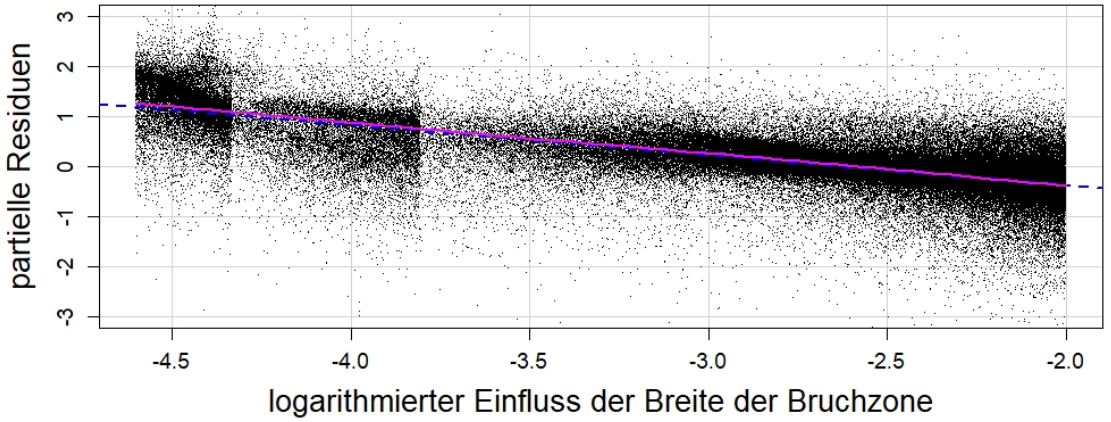

Abbildung 15: partielle Residuen, um den Haupteffekt von der Bruchzonenbreite *hwid* bereinigt.

Zum Anderen fällt in Abbildung 13 auf, dass die Residuen für die logarithmierten, geschätzte Werte $\widehat{error}_{p_i}$ der relativen Abweichung zwischen numerischer und analy-

tischer Lösung bezüglich des Porendrucks nahe des Werts −7 eine größere Streuung besitzen, als bei den geschätzten Werten auf dem restlichen Wertebereich. Dabei ist jedoch zu beachten, dass die meisten der geschätzten Werte auch in genau jenem Bereich liegen (siehe Abbildung 16).

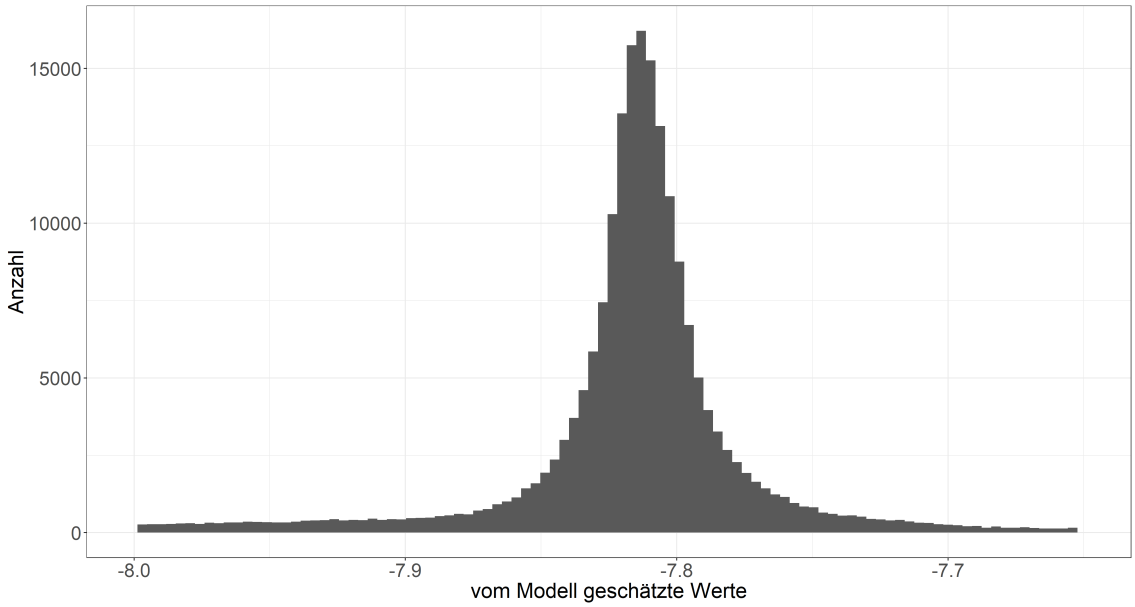

Abbildung 16: Vom Modell geschätzte Werte und die jeweilige Anzahl

Da also der Großteil der geschätzten Werte (über 78%) im Intervall $[-8, -7.5]$ liegt, ist mit mehr absoluter Streuung zu rechnen. Um dem entgegenzuwirken, wird beim Invertieren des Modells die Verteilung der inversen Schätzung über *Bootstrap* geschätzt und entsprechend ein Wert gesucht, welcher mit einer Wahrscheinlichkeit von über 95% die Fehlerschranke 1.0E−06 nicht überschreitet. Dies ist in dem Fall der Randpunkt des einseitigen 95%-Konfidenzintervalls.

4.4 Inversion

4.4.1 Schätzung der numerischen Parameter

Im Methodikteil dieser Abschlussarbeit wird in Kapitel 3.6 erläutert, auf welche Art und Weise $X_0 \in [X_{0min}, X_{0max}]$ ermittelt wird. $X_0 \in [X_{0min}, X_{0max}]$ bzw. in diesem Fall $nz_0 \in [10, 1000]$ wird hierbei, da es sich bei nz um Ganzzahlen handelt, zur Menge $X_0 \in [X_{0min}, X_{0max}] \cap N \Leftrightarrow X_0 \in [X_{0min}, .., X_{0max}]$. Diese Berechnungen wurden mit dem R-Package `investr` (Greenwell and Kabban; 2014) gemacht. Da dieses stetige Intervalle durchsucht, wurde dies dadurch modifiziert, dass die jeweils berechnete Zahl $nz \in [10, 1000]$ abgerundet wurde. Das R `investr`-

Package (Greenwell and Kabban; 2014) bietet auch gleichzeitig eine, hier genutzte, Implementierung zur Bestimmung von entsprechenden *Bootstrap*-Intervallen. Dabei wurden einige Funktionen für die in diesem Fall benötigte Verwendung minimal modifiziert, aber die grundsätzlichen Berechnungsfunktionen wurden beibehalten und verwendet. Die Errechnung der Anzahl an Gridpunkten (nz) ergibt sich also zu:

$$\begin{aligned}(1.0\text{E}{-}06) - \hat{f}(\alpha_{th}, \alpha_{hy}, \lambda, hwid, nz) \stackrel{!}{=} 0,\\ \text{für } nz_0 \in [10, ..., 1000]\end{aligned} \tag{28}$$

Die entsprechenden approximativen Prognoseintervalle werden mittels *Bootstrap* geschätzt. Dafür werden aus dem Datensatz, mit welchem das Modell geschätzt wurde $n_{sim} = 100$ *Bootstrap* $\boldsymbol{X}_1, ..., \boldsymbol{X}_{100}$, Stichproben gezogen. Für jede dieser Stichproben werden die Parameter ($\boldsymbol{\Theta}$) des entwickelten Modells aus Abschnitt 4.3 neu geschätzt. Aus den daraus entstehenden Schätzungen der $\hat{\boldsymbol{\beta}}_{i,j}, \hat{\boldsymbol{\gamma}}_{i,j} \in \hat{\boldsymbol{\Theta}}_j$ mit $j \in 1, ..., n_{sim}$ wird jeweils, wie bereits beschrieben, $nz_{0,j}$ geschätzt. Aus eben jenen Schätzungen $\hat{nz}_{0,j}$ wird durch deren empirische Verteilung und des geschätzten Standardfehlers $SD_{nz0,j}$ ein Intervall konstruiert, welches wie in Abschnitt 3.6.2 errechnet wird. Gleichzeitig bestimmt man eine Punktschätzung für nz_0, für welche mit 95-prozentiger Wahrscheinlichkeit die gegebene Fehlerschranke 1.00E−06 nicht überschritten wird. Diese abschließende Schätzung ergibt sich also zu:

$$\hat{nz}_0 = [\overline{nz_{0,j}} + 1.64\frac{\hat{SD}_{nz_0}}{\sqrt{n_{sim}}}] \tag{29}$$

Für nz_0 wurde hier der Mittelwert der n_{sim}-Bootstrapschätzungen für $nz_{0,j}$ des `investr`-Packages genutzt, welche wie in Abschnitt 3.6 erläutert berechnet wurde. Fälle für welche die Schätzung von nz_0 über 1000 sein würde wurden ignoriert, da ein Regressionsmodell nur über den, für die Modellierung benutzten Parameterintervallen, stabil ist.

4.4.2 Validierung

Um die Punkt- sowie die *Bootstrap*-Schätzungen zu validieren, werden für eben jene in dieser Abschlussarbeit berechneten Schätzungen für nz_0 die Simulationsberechnungen durchgeführt, welche die relativen Abweichungen zwischen analytischer und numerischer Lösung bezüglich des Porendrucks errechnen. Anhand dessen kann beurteilt werden, wie gut die verwendete Modellierung und die zugehörige Inversion funktioniert hat. Dementsprechend wird betrachtet, zu welcher relativen Abweichung zwischen analytischer und numerischer Lösung bezüglich des Porendrucks die

Schätzungen führten. Dabei gilt es, einen guten Kompromiss zwischen Genauigkeit und Rechenzeit zu finden. Hierbei sollten 95% der Schätzungen der relativen Abweichung zwischen analytischer und numerischer Lösung bezüglich des Porendrucks unter der gegebenen Fehlerschranke von 1.00E−06 liegen. Dies wäre aber beispielsweise auch zu erreichen, wenn man die Anzahl der Gridpunkte auf ihren Maximalwert $nz = 1000$ setzen würde. Um jedoch eine, bezüglich der Rechenzeit optimale, Lösung zu finden, ist der Anspruch an die Schätzungen eben jene 95% zu erreichen, aber den numerischen Parameter der Gridpunktanzahl nicht systematisch zu überschätzen. Von den 87 Schätzungen für nz_0 haben 76 die Fehlerschranke nicht überschritten.

5 Fazit und Ausblick

Aufbauend auf der in Abschnitt 3 erläuterten Methodik und deren Anwendung in Abschnitt 4 lassen sich die numerischen Parameter aus gegebenen physikalischen Fehlern berechnen. Dabei wurde erörtert, aus welchem Grund es sinnvoll ist, den Parameter Dwn_{min} auf seinen Minimalwert $\frac{1.00\text{E}{-16}}{hwid}$ zu setzen. Durch diese deterministische Abhängigkeit von einem physikalischen Parameter ergibt sich die Problemstellung, durch die ausschließliche Bestimmung eines geeignetes Werts für die Anzahl an Gridpunkten, neu. Diese kann anhand einer Inversion einer Regression bestimmt werden. Das Modell erzielte dabei zu 87% zufriedenstellende Ergebnisse. Abschließend ist anzumerken, dass bei der Modellierung der relativen Abweichung systematische Fehler auftauchten, welche in Abschnitt 4.3 aufgezeigt wurden. Hier könnte eine noch gezieltere Modellierung der Breite der Bruchzone helfen. Beispielsweise wäre dort eine Modellierung durch Splines mit mehr Knoten möglich, was auf Grund der Interaktionen im Rahmen dieser Abschlussarbeit nicht ausgewertet werden konnte, da die benutzten Computer für eine solche Berechnung bei der Datenmenge nicht ausreichten. Auch eine zweistufige gewichtete Regression kommt bei Varianzheteroskedastizität in Frage. Eine solche konnte die vorliegenden Probleme jedoch nicht beheben (siehe Plot im Anhang in Abschnitt 5). Außerdem wurden für die meisten der Schätzungen aufgrund der Rechenintensität nur $n_{sim} = 100$ *Bootstrap*-Zufallsstichproben gezogen. Die Erhöhung davon sollte zu noch genaueren Ergebnissen führen. Die Modellierung errechnet trotz dieser kleineren Schwächen in 82 von 96 Berechnungen einen Wert für nz_0, welcher dazu führt, dass die Simulation die Fehlerschranke nicht überschreitet. Durch die in dieser Abschlussarbeit erstellten Modelle, konnten enstprechend die für eine stabile Berechnung benötigen numerischen Parameter anhand von physikalischen bestimmt werden. Dieses Modell

sollte nun nutzbar sein, um die Erdbebensimulationen in Bezug auf ihre Rechenzeit zu optimieren, ohne dabei die nötige Präzision zu verlieren.

Literatur

Alfons, A. (2012). *cvTools: Cross-validation tools for regression models*. R package version 0.3.2.
URL: *https://CRAN.R-project.org/package=cvTools*

Auguie, B. (2017). *gridExtra: Miscellaneous Functions for „Grid“ Graphics*. R package version 2.3.
URL: *https://CRAN.R-project.org/package=gridExtra*

Becker, Claudia und Genschel, U. (2005). *Schließende Statistik Grundlegende Methoden*, Springer-Verlag Berlin Heidelberg.

Bhattacharya, Rabi und Lin, L. u. P. V. (2016). *Ein Kurs in mathematischer Statistik und großer Stichprobentheorie*, 1st edn, Springer Publishing Company, Incorporated.

Blobel, V. and Lohrmann, E. (2013). *Statistische und numerische Methoden der Datenanalyse*, Teubner Studienbücher Physik, Vieweg+Teubner Verlag.

Boos, D. D. u. S. L. A. (2013). *Essential Statistical Inference*, 1st edn, Springer-Verlag New York.

Draper, N. R. and Smith, H. (2014). *Fitting Straight Lines: Special Topics*, John Wiley Sons, Ltd, chapter 3, pp. 79–114.

Fahrmeir, L., Kneib, T. and Lang, S. (2009). *Regression. Modelle, Methoden und Anwendungen*, 2 edn, Springer.

Fox, J. and Weisberg, S. (2019). *An R Companion to Applied Regression*, third edn, Sage, Thousand Oaks CA.
URL: *https://socialsciences.mcmaster.ca/jfox/Books/Companion/*

from Jed Wing, M. K. C., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C. and Hunt., T. (2019). caret: Classification and regression training. R package version 6.0-84.
URL: *https://CRAN.R-project.org/package=caret*

Gramacy, R. B. and Taddy, M. (2010). Categorical inputs, sensitivity analysis, optimization and importance tempering with tgp version 2, an R package for treed gaussian process models, *Journal of Statistical Software* **33**(6): 1–48.
URL: *http://www.jstatsoft.org/v33/i06/*

Greenwell, B. M. and Kabban, C. M. S. (2014). investr: An r package for inverse estimation, *The R Journal* **6**(1): 90–100.
URL: *http://journal.r-project.org/archive/2014-1/greenwell-kabban.pdf*

James, Gareth und Witten, D. u. H. T. u. T. R. (2014). *Eine Einführung in das statistische Lernen: Mit Anwendungen in R*, Springer Publishing Company, Incorporated.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Scatterplot Smoothing*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, p. 57–90.

Stein, M. (1987). Large sample properties of simulations using latin hypercube sampling, *Technometrics* **29**(2): 143–151.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.
URL: *https://ggplot2.tidyverse.org*

Abbildungsverzeichnis

Tabellenverzeichnis

Eigenständigkeitserklärung

Hiermit bestätige ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Quellen entnommen sind, wurden unter Angabe der Herkunft kenntlich gemacht. Hierzu zählen auch Zeichnungen, Skizzen sowie Internetquellen.

Ort, Datum Unterschrift

Anhang

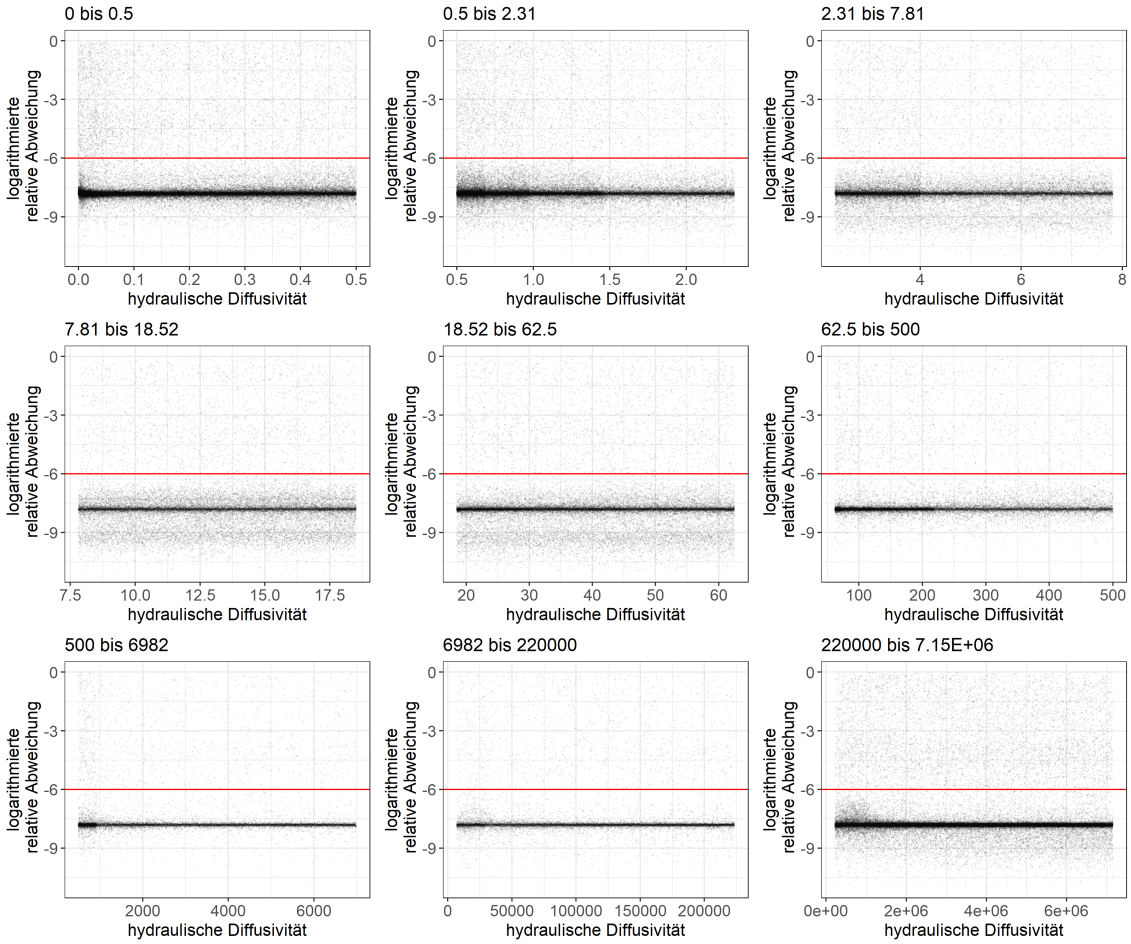

Abbildung 17: Einfluss der hydraulischen Diffusivität auf die relative Abweichung zwischen numerischer und analytischer Lösung bezüglich des Porendrucks.

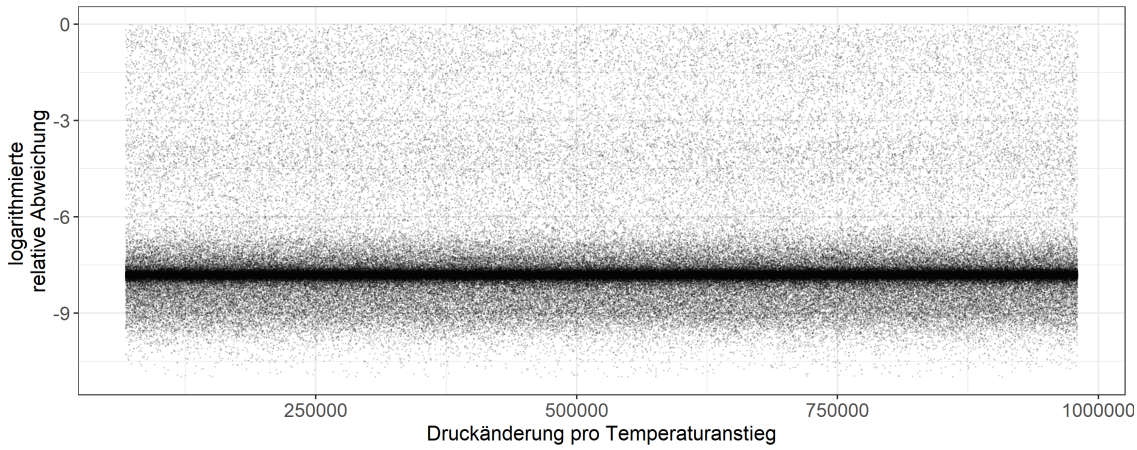

Abbildung 18: Einfluss der Druckänderung pro Temperaturanstieg λ auf die logarithmierte relative Abweichung zwischen numerischer und analytischer Lösung bezüglich des Porendrucks

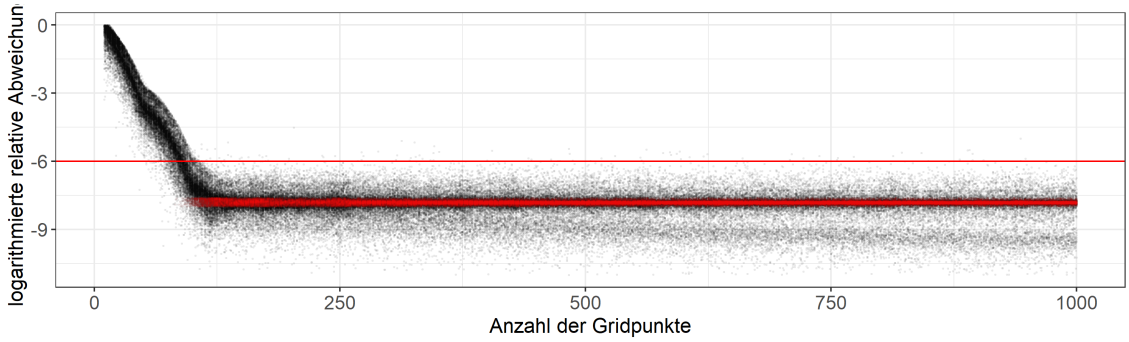

Abbildung 19: Die vom Modell fehlspezifizierten Datenpunkte liegen scheinbar alle im Bereich Anzahl der Gridpunkte $nz > 100$.

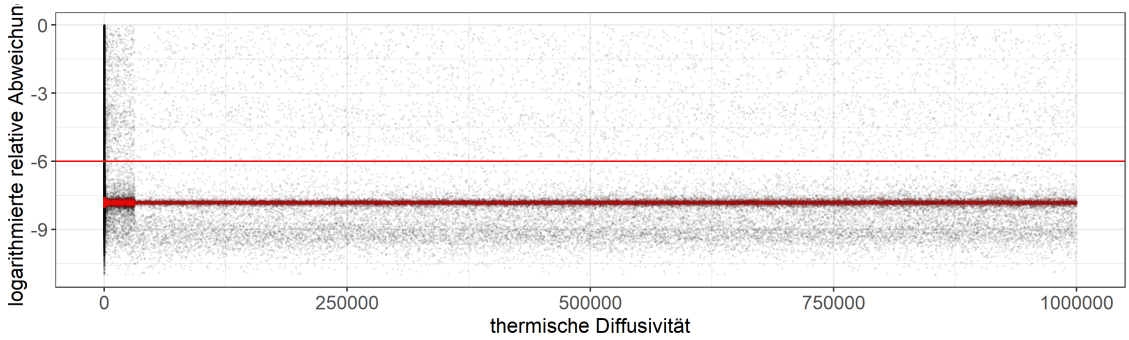

Abbildung 20: Der Einfluss der thermischen Diffusivität scheint keine Bereiche zu beinhalten in welchem mehr oder weniger systematische Unter- bzw. Überschätzungen liegen

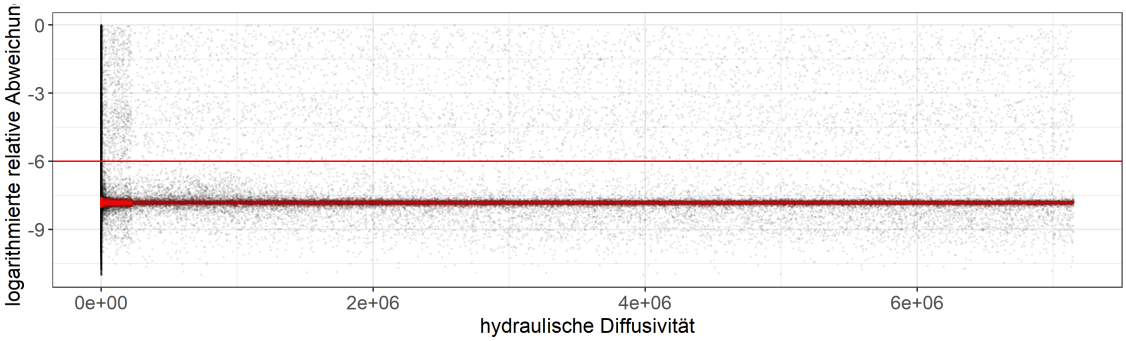

Abbildung 21: Der Einfluss der hydraulischen Diffusivität scheint keine Bereiche zu beinhalten in welchem mehr oder weniger systematische Unter- bzw. Überschätzungen liegen

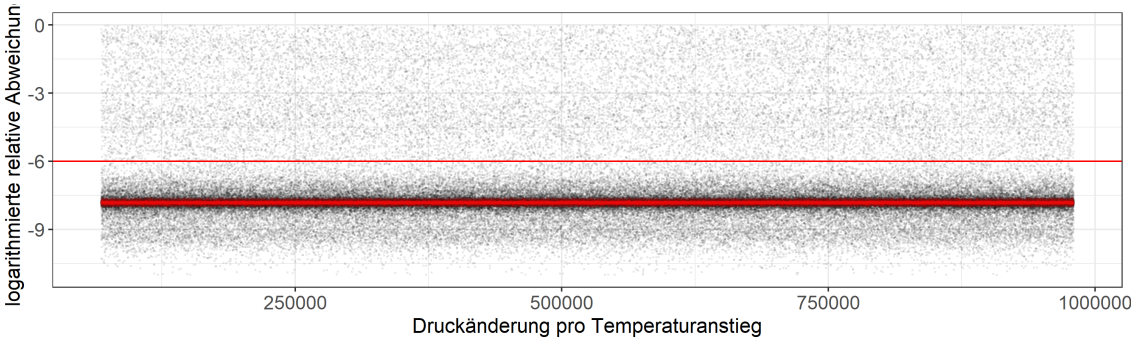

Abbildung 22: Der Einfluss der Druckänderung pro Temperaturanstieg scheint keine Bereiche zu beinhalten in welchem mehr oder weniger systematische Unter- bzw. Überschätzungen liegen

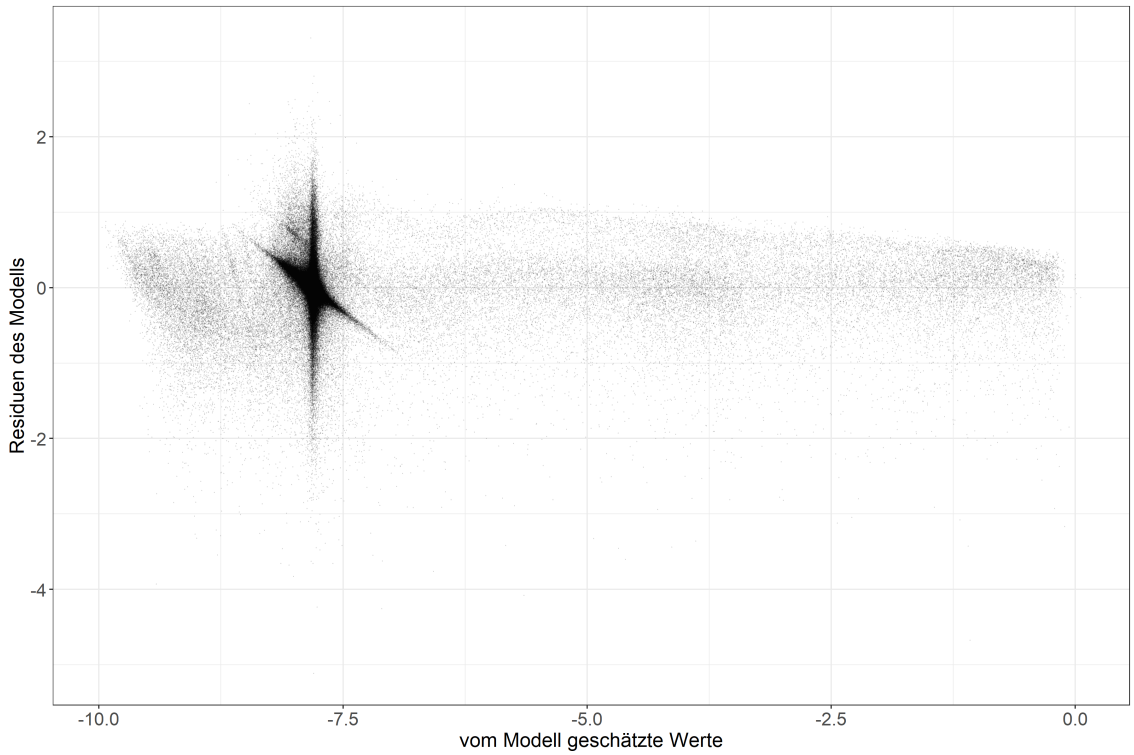

Abbildung 23: Residuenverteilung bei einer zweistufigen Schätzung nach Fahrmeir et al.; 2009 (S. 133). Die Gewichte wurden über eine Regression auf die quadrierten Residuen des in Abschnitt 4.3 verwendeten Modells bestimmt. Die Gewichte $\hat{\omega}$ ergaben sich zu $\hat{\omega} = \frac{1}{exp(\hat{\eta})}$ mit $\hat{\eta} = \hat{\epsilon}^2$.

Die in R genutzten Codes befinden sich, wie mit dem Betreuer abgesprochen in einem Git-Repsitory mit dem Commit „#Abgabe“.